MASTER EN SCIENCES MATHÉMATIQUES

Recherche de points fixes au moyen d'une homotopie

Henkes, Oswald

Award date:
1980

Awarding institution:
Universite de Namur

Année scolaire 1979-80

RECHERCHE DE POINTS FIXES AU MOYEN D'UNE HOMOTOPIE

Oswald Henkes

To be is to do (Sartre)

To do is to be (Nietzsche)

Do be do be do (Sinatra)

Je remercie bien vivement Monsieur Nguyen van Tien d'avoir accepté la direction de ce mémoire.

Je tiens aussi à exprimer toute ma reconnaissance à Monsieur Jean-Jacques Strodiot pour l'aide efficace apportée au long de ce travail.

Enfin, je remercie tous ceux qui de près ou de loin ont contribué à la réalisation de ce mémoire.

Introduction

Le but de ce travail est de présenter une méthode efficace pour la recherche des points fixes d'une application régulière f de $\mathbb{R}^n$ dans $\mathbb{R}^n$. Cette méthode due à Chow, York et Mallet [4] est basée sur une démonstration constructive du théorème du point fixe de Brouwer faisant intervenir des arguments de géométrie différentielle.

Remarquons que chercher $\bar{x}$ tel que $f(\bar{x}) = \bar{x}$ revient à résoudre le système $f(x) - x = 0$. La méthode consiste alors à considérer une fonction très simple $g_a(x)$ dont $a \in \mathbb{R}$ est une racine (en général, on choisit $g_a(x) = x - a$) et ensuite à la transformer de manière régulière de telle façon que les racines de la fonction transformée tendent vers

un point fixe de f.

Plus précisément on définit l'application homotopique

$$\Phi_a(\lambda, x) = (1-\lambda)\, g_a(x) + \lambda\,(x - f(x)) \qquad , 0 \leq \lambda \leq 1$$

qui transforme $g_a(x)$ en $x - f(x)$ lorsque λ parcourt l'intervalle $[0,1]$ et on considère la partie $\Phi_a^{-1}(0)$ de $[0,1] \times \mathbb{R}^n$. On montre alors que pour presque tout $a \in \mathbb{R}^n$ la composante de $\Phi_a^{-1}(0)$ qui contient a est une courbe régulière Γ_a de longueur finie joignant le point $(0,a)$ à un point $(1,\bar{x})$ où $\bar{x}$ est un point fixe cherché (voir fig. 1).

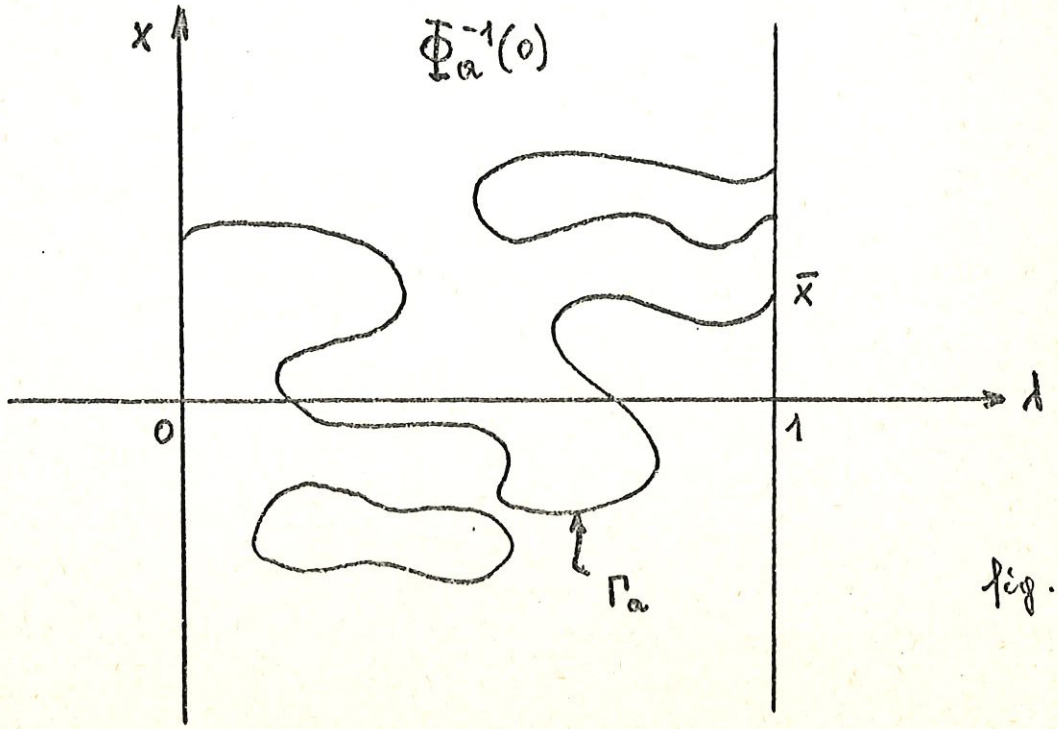


fig. 1

La méthode consistera alors à suivre la courbe Γ_a en partant de $(0,a)$ ce qui pourra se faire en résolvant l'équation différentielle

$$\begin{cases} \dfrac{d\,\Phi_a(y)}{dy} \cdot \dfrac{dy}{ds} = 0 \\ y(0) = (0,a) \end{cases}$$

où $y = (\lambda, x)$ et s est la longueur d'arc de courbe.

Après avoir établi rigoureusement les bases de la méthode de Chow, Yorke et Hallet, le travail a surtout consisté à mettre au point un code efficace et à le tester sur des problèmes de programmation mathématique avec et sans contraintes, sur des problèmes de complémentarité, sur des problèmes aux limites et sur des problèmes de recherche de racines de polynômes.

Les chapitres I et II sont consacrés au théorème de Brouwer, le chapitre III décrit la méthode de Chow, Yorke et Hallet de façon détaillée et les trois derniers chapitres sont consacrés aux applications.

CHAPITRE I :

Le théorème du point fixe de Brouwer :

Une preuve non constructive

Nous allons établir dans ce chapitre une démonstration non constructive du théorème du point fixe de Brouwer. Ceci sera fait de façon assez rapide pour montrer un aspect différent à celui du chapitre II. Les approches utilisées seront surtout du type topologique et se basent sur [1].

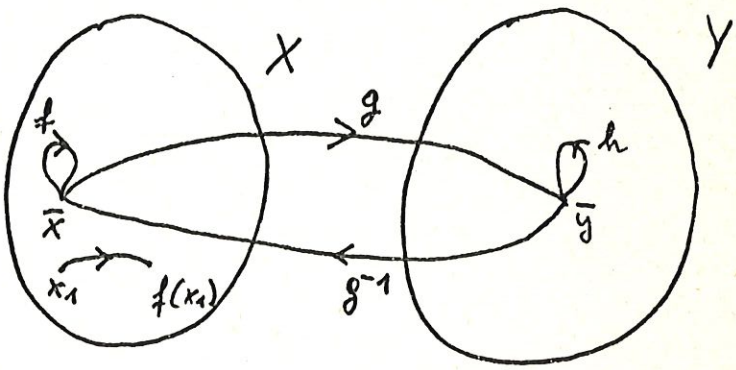
Définition I.1. : Un espace topologique X admet la propriété du point fixe ssi toute application continue de X dans X admet un point fixe.

Il est souvent possible de décider qu'un ensemble n'a pas la propriété du point fixe en trouvant une application continue sans points fixes. Ceci est le cas pour la droite réelle (en considérant $f(x) = x + 1$) et le cercle unité (en

utilisant $f(x) = -x$). Par contre l'intervalle $[0;1]$ admet bien la propriété du point fixe : Toute application continue f de $[0;1]$ vers $[0;1]$ possède un point $\bar{x}$ tel que $f(\bar{x}) = \bar{x}$.

Théorème I.2. : Si X est homéomorphe à Y et X admet la propriété du point fixe, alors il en sera même pour Y.

Preuve : Soit g un homéomorphisme de X vers Y. On a donc que g et g^{-1} sont continues. Soit $\bar{x}$ le point fixe de f sur X et soit $\bar{y} = g(\bar{x})$,



f étant une application continue quelconque. Comme en plus tout h de Y vers Y peut s'exprimer comme $h = g \circ f \circ g^{-1}$ pour un certain f, on déduit que $g(f(g^{-1}(\bar{y}))) = g(f(\bar{x})) = g(\bar{x}) = y$. Par conséquent $\bar{y}$ est le point fixe de l'application continue h. ∎

Définition I.3. : L'espace topologique X est une rétraction de Y ssi $X \subset Y$ et il existe une application continue r de Y vers X telle que $r = id$ sur X.

Dans ce cas r est appelé application rétractive.

Théorème I.4. : Si Y a la propriété du point fixe
X est une rétraction de Y
alors X a la propriété du point fixe.

Preuve: Soit r la rétraction de Y vers X. Si T est une application continue quelconque de X vers X, alors $T \circ r$ est une application continue de Y vers X. Comme $T \circ r$ va de Y dans Y, il existe par hypothèse un point fixe w de Y tel que $(T \circ r)(w) = w$. Comme $w \in X$ et $r = id$ sur X, on a que $rw = w$ et finalement $Tw = w$. ∎

Dans la suite quelques notions de contractibilité nous seront très utiles. Soit $\overline{B}^n = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\}$ et $S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$.

Définition I.5. :

Un espace topologique X est contractible (à un point x_0 de X)
ssi il existe une fonction continue $f(x,t)$ de $X \times [0,1]$ vers X
telle que $f(x,0) = x$ et $f(x,1) = x_0$ pour tout $x \in X$
(f est appelé contraction de X à x_0).

Remarques I.6 :

I.6.1. On a que tout convexe X de $\mathbb{R}^n$, $n \geq 1$, est contractible. Soit pour cela un $x_0 \in X$ et $h : X \times [0,1] \longrightarrow X$ tel que (x,t) soit envoyé sur $h(x,t) \equiv (1-t)x + t x_0$. Tout point de X est donc envoyé sur le segment entre lui-même et x_0, ce qui est bien permis par convexité de X.

Il en suit évidemment que $\overline{B}^n$ est contractible.

I.6.2. Pour $n \geq 1$, S^{n-1} est non contractible. Dans le cas $n=1$, S^{n-1} se réduit à $\{1,-1\}$. Une contraction doit être de la forme $h(x,t) = (1-t)x + t x_0$. En prenant p.ex. $x_0 = 1$, on a que $h(-1, 1/3) = -1/3 \notin S^1$. Il n'y a pas moyen de trouver une contraction envoyant S^0 sur 1 ou -1. Pour le cas n quelconque on peut se baser sur les réflexions de [1] et [2].

Lemme I.7. : Si Y est contractible, alors toute rétraction X de Y l'est.

Preuve: Soit r l'application rétractive de Y vers X et soit $f(x,t)$ la fonction qui contracte Y à un point $z \in Y$. Comme $r = id$ sur X, il suit que $r \circ f(x,t)$ contracte X à un point $r(z) \in X$. ∎

Théorème I.8. : Pour $n \geq 1$, S^{n-1} n'est pas une rétraction de $\overline{B}^n$.

Preuve: $\overline{B}^n$ est contractible ; comme S^{n-1} ne l'est pas, le résultat suit du lemme précédent.

Théorème I.9. : Théorème du point fixe de Brouwer (1910).

(i) $\overline{B}^n$ admet la propriété du point fixe

(ii) Tout ensemble compact, connexe et non-vide X de $\mathbb{R}^n$ admet la propriété du point fixe.

Preuve: (i) Si par contradiction, il existe une application continue T de $\bar{B}^n$ vers $\bar{B}^n$ sans point fixe, alors définissons l'application rétractive r de $\bar{B}^n$ vers S^{n-1} de la façon suivante. Pour tout $x \in \bar{B}^n$, on va étendre le segment de droite de Tx vers x vers un point d'intersection avec S^{n-1} (voir fig 1). Appelons ce point rx. Or, une telle application rétraction est impossible par le thm I.8.

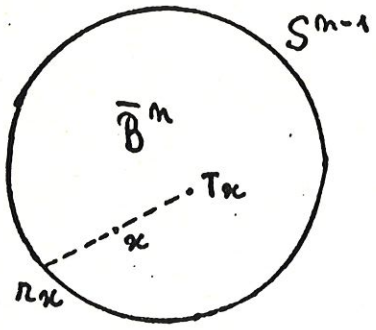


fig. 1

$$S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$$
$$\bar{B}^n = \{x \in \mathbb{R}^n \mid \|x\| \leq 1\}$$
$$B^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$$

(ii) Pour k suffisamment grand, la boule $k \cdot \bar{B}^n$ de rayon k contient X. Comme un convexe fermé non-vide X de E^n est une rétraction de tout sous-ensemble de E^n qui le contient (voir [3]), on a que X est une rétraction de $k\bar{B}^n$. Comme $k \cdot \bar{B}^n$ est homéomorphe à $\bar{B}^n$, le thm I.2 montre que $k \cdot \bar{B}^n$ admet la propriété du point fixe. Le thm I.4. entraîne alors que X admet la propriété du point fixe. ∎

Le thm précédent analyse la propriété du point fixe non seulement au cas de $\bar{B}^n$, mais pour tout compact convexe de $\mathbb{R}^n$ ce qui représente un gain de généralité.

CHAPITRE II :

Le théorème du point fixe de Brouwer :

Une preuve constructive

Le chapitre fournira une preuve du théorème du point fixe de Brouwer tout en construisant un algorithme pour la recherche de ces points fixes. Cet algorithme est caractérisé par le théorème II.3.2. qui est par conséquent le résultat fondamental de ce paragraphe. Les éléments théoriques seront obtenus surtout à partir du théorème de Sard.
Nous suivons en gros le raisonnement de Chow, York et Yorke [4].

Soit donné une fonction régulière f (c-à-d de classe $C^2, C^p, p>2$ ou C^∞) de $\bar{B}^n$ vers $\bar{B}^n$ dont on cherche les points fixes.
Utilisons dans la suite en plus une fonction régulière g_a (a étant un paramètre appartenant à $\mathbb{R}^n$) de $\bar{B}^n$ vers $\bar{B}^n$. Nous allons

construire une application homotopique Φ_a de $\overline{B}^n \times [0,1]$ vers $\overline{B}^n$ qui doit par définition satisfaire la condition suivante

$$\Phi_a(x,1) = x - f(x)$$
$$\Phi_a(x,0) = g_a(x) \qquad \text{pour tout } x \in \overline{B}^n$$

Dans ce cas on dit que Φ_a est une application homotopique (ou homotopie) de $g_a(x)$ et $x - f(x)$.

La condition est vérifiée pour $\Phi_a(x,\lambda) = \lambda(x - f(x)) + (1-\lambda)\, g_a(x)$.

Le paramètre $a \in \overline{B}^n$ étant arbitraire, on peut considérer l'application $\Phi(a,\lambda,x) \equiv \Phi_a(\lambda,x)$ où cette fois-ci Φ envoit $\overline{B}^n \times [0,1] \times \overline{B}^n$ sur $\overline{B}^n$. $\Phi(a,\lambda,x)$ est évidemment régulière.

II.1. Le théorème de Sard

Le théorème de Sard fournit une base essentielle de la construction de l'algorithme de recherche des points fixes de f. Rappelons d'abord quelques définitions de la géométrie différentielle. Les définitions ainsi que le théorème de Sard sont donnés dans [5] et [6].

II.1.1. Définitions

1. Un sous-espace $M \subset \mathbb{R}^n$ est appelé <u>surface régulière</u> de dimension m ssi chaque $x \in M$ admet un voisinage $W \cap M$ qui est difféomorphe à un ouvert de l'espace euclidien $\mathbb{R}^m$.

Une courbe régulière par exemple est un sous-espace qui admet en chacun de ses points un voisinage difféomorphe à un ouvert de $\mathbb{R}$. Le fait que l'application "identité" est un difféomorphisme implique que tout ouvert de $\mathbb{R}$ est une courbe régulière.

La sphère unité S^2 étant les points $(x,y,z) \in \mathbb{R}^3$ tels que $x^2+y^2+z^2=1$ est un sous-espace régulier de dimension 2. Considérons en effet le difféomorphisme envoyant (x,y) sur $(x,y,\sqrt{1-x^2-y^2})$ pour $x^2+y^2<1$. Il paramétrise évidemment la région $z>0$ de S^2. En changeant signe et rôle des variables on observe une paramétrisation des régions $x>0, y<0, x<0, y>0$ et $z<0$ qui sont tous des ouverts de $\mathbb{R}^2$. Comme ces régions recouvrent S^2, on conclut que S^2 est un sous-espace régulier de dimension 2.

2. Considérons U un ouvert de $\mathbb{R}^n$ et f une fonction régulière de U dans $\mathbb{R}^p$ avec $n \geq p$.

Alors, $y \in \mathbb{R}^p$ est <u>valeur régulière</u> de f

ssi le rang de $Df(x)$ (noté $rg(Df(x))$) vaut p pour tout $x \in f^{-1}(y)$ où $Df(x)$ représente la matrice jacobienne de f en x.

L'ensemble $C \equiv \{ x \in U \mid rg(Df(x)) < p \}$ sera appelé l'ensemble des <u>points critiques</u> de f et $f(C)$ l'ensemble des <u>valeurs critiques</u> de f.

Il résulte de ces définitions que $\mathbb{R}^p \setminus f(C)$ représente l'ensemble des valeurs régulières de f.

II.1.2. Le théorème de Sard

Soit $f : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}^p$ régulière

Alors, ① Presque tout $y \in \mathbb{R}^p$ est valeur régulière de f

② Si $C \equiv \{ x \in U \mid rg(Df(x)) < p \ ; x = f^{-1}(y) \}$, alors $f(C)$ est de mesure nulle dans $\mathbb{R}^p$.

Par définition les thèses ① et ② sont équivalentes.

Preuve: Nous allons procéder par récurrence sur la dimension n.

→ Comme $\mathbb{R}^0$ représente un seul point, le théorème sera certainement vrai pour $n=0$. Ceci représente le début de la récurrence.

Soit $C_1 \equiv \{ x \in U \mid$ la dérivée première de f en x est nulle $\}$.

$C_i \equiv \{ x \in U \mid$ toutes les dérivées partielles de f d'ordre plus petit que i s'annullent en $x \}$.

On aura par conséquent $C \supset C_1 \supset C_2 \supset \dots$

Comme $C = (C \setminus C_1) \cup (C_1 \setminus C_2) \cup \dots \cup (C_i \setminus C_{i+1}) \cup \dots \cup C_k$ où k grand il suffit de démontrer que

#1 : L'image $f(C \setminus C_1)$ est de mesure nulle.

#2 : L'image $f(C_i \setminus C_{i+1})$ est de mesure nulle pour $i \geq 1$

#3 : L'image $f(C_k)$ est de mesure nulle pour k assez grand

Il se peut que l'intersection des C_i soit vide si f est constant sur une composante entière de U. Ce cas est suffisant pour exclure #1 et #2.

→ On suppose donc les trois pas satisfaits pour $n-1$

→ Voyons ce qui se passe pour n :

#1 : On peut évidemment supposer $p \geq 2$ car $C = C_1$ si $p = 1$.

Nous utilisons dans la suite <u>le théorème de Fubini</u> (voir [7]) :

> Un ensemble mesurable $A \subset \mathbb{R}^p = \mathbb{R} \times \mathbb{R}^{p-1}$ est de mesure nulle s'il coupe tout hyperplan (constante) $\times \mathbb{R}^{p-1}$ en un ensemble de mesure nulle (de dimension $p-1$).

Pour tout $\bar{x} \in C - C_1$ on peut trouver un voisinage ouvert $V \subset \mathbb{R}^n$ tel que $f(V \cap C)$ est de mesure nulle (1). Comme $C - C_1$ est couvert par un nombre au plus dénombrable de ces voisinages, ceci prouvera que $f(C - C_1)$ est de mesure nulle. Reste donc à démontrer (1).

Comme $\bar{x} \notin C_1$, il y a une dérivée partielle soit p. ex. $\frac{\partial f_1}{\partial x_1}$ non nulle en $\bar{x}$. Considérons l'application $h : U \longrightarrow \mathbb{R}^n$ définie par $h(x) = (f_1(x), x_2, x_3, \dots, x_n)$. Il en suit que

$$Dh(\bar{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x}) & 0 \\ \hline * & \begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix} \end{pmatrix}$$

et donc que $Dh(\bar{x})$ est non-singulière.

Il en suit que h transporte un voisinage V de $\bar{x}$ au moyen d'un difféomorphisme dans un ouvert V' de $\mathbb{R}^n$ (par le thm des fonctions implicites (thm II.1.4.)). La composition $g = f \circ h^{-1}$ transfère V' dans $\mathbb{R}^p$ et $f = g \circ h$ (voir fig. 1).

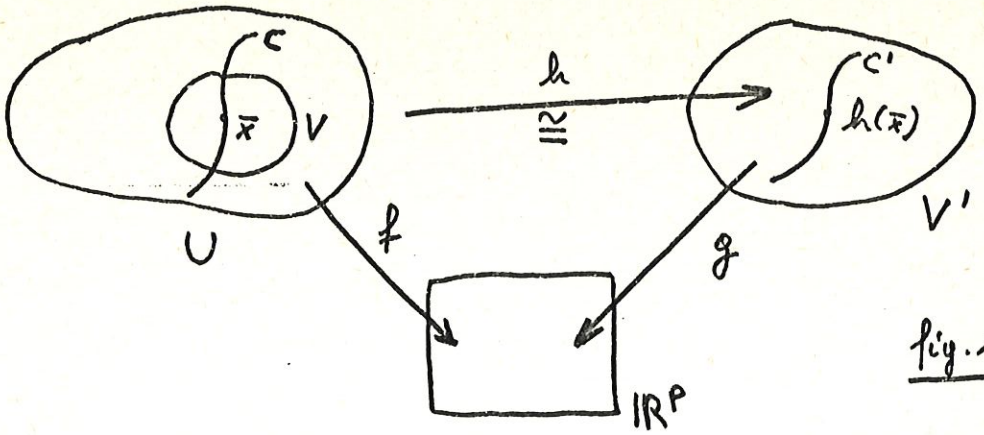


fig. 1

Définissons $C' \equiv \{$ points critiques de $g\}$. On aura donc que $x \in C'$ $\Leftrightarrow rg(Dg(x)) < p \Leftrightarrow rg\left[D(f \circ h^{-1})(x)\right] < p \Leftrightarrow rg\left[f(f_1^{-1}(x), x_2 \dots x_n)\right] < p$

$\Leftrightarrow x' \in V \cap C$ où $x_1' = f_1^{-1}(x)$ et $x' = (x_1', x_2, \dots, x_n)$

$\Leftrightarrow x \in h(V \cap C)$. Il suit que $C' = h(V \cap C)$ et par conséquent que $g(C') = f(V \cap C) \equiv$ ensemble des valeurs critiques de g.

Pour tout $(t, x_2, \dots, x_n) \in V'$ remarquons que $g(t, x_2 \dots x_n) = f(x_1 \dots x_n)$ appartient à l'hyperplan $t \times \mathbb{R}^{p-1} \subset \mathbb{R}^p$: d'où, g envoie hyperplan sur hyperplan. Soit $g' : (t \times \mathbb{R}^{n-1}) \cap V' \longrightarrow t \times \mathbb{R}^{p-1}$ la restriction de g sur V'. La matrice des dérivées premières de g est de la forme $\left(\frac{dg_i}{dx_i}\right) = \begin{bmatrix} 1 & 0 \\ * & dg_i'/dx_i \end{bmatrix}$. Donc, un point de $t \times \mathbb{R}^{n-1}$ est point critique pour g' ssi il l'est pour g.

Par hypothèse de récurrence, l'ensemble des valeurs critiques de g' est de mesure nulle dans $t \times \mathbb{R}^{p-1}$. Donc, l'ensemble des valeurs critiques de g coupe tout hyperplan $t \times \mathbb{R}^{p-1}$ en un ensemble de

mesure nulle (par déf. de g). Cet ensemble $g(C')$ est mesurable, car il peut être exprimé comme union au plus dénombrable de sous-ensembles compacts (comme étant une partie de $\mathbb{R}^p$).

Par le thm de Fubini on a que $g(C')$ et par conséquent $f(V \cap C)$ est de mesure nulle. Ceci démontre (*) et achève donc #1.

#2 : Pour tout $\bar{x} \in C_k \setminus C_{k+1}$ il y a une des $k+1^{\text{èmes}}$ dérivées $\dfrac{\delta^{k+1} f_r}{\delta x_{s_1} \dots \delta x_{s_{k+1}}}$ qui n'est pas nulle. La fonction $w(x) = \dfrac{\delta^k f_r}{\delta x_{s_2} \dots \delta x_{s_{k+1}}}$ s'annulle en $\bar{x}$ sans que $\dfrac{\delta w(\bar{x})}{\delta x_{s_1}}$ soit zéro. Alors, $h : U \to \mathbb{R}^n$ définie par $h(x) = (w(x), x_2 \dots x_n)$ est telle que $Dh(\bar{x})$ est non-singulier. h envoit donc le voisinage V de $\bar{x}$ dans un ouvert V' de $\mathbb{R}^n$ par un difféomorphisme. Notons que h envoit $C_k \cap V$ sur l'hyperplan $0 \times \mathbb{R}^{n-1}$. Considérons de nouveau $g = f \circ h^{-1} : V' \longrightarrow \mathbb{R}^p$ c-à-d $f = g \circ h$. Soit $\bar{g} : (0 \times \mathbb{R}^{n-1}) \cap V' \longrightarrow \mathbb{R}^p$ la restriction de g. Par hypothèse de récurrence, l'ensemble des valeurs critiques de $\bar{g}$ est de mesure nulle dans $\mathbb{R}^p$. Or, l'ensemble des points critiques de g et $\bar{g}$ est le même (#1). En plus, tout point dans $h(C_k \cap V)$ est point critique de $\bar{g}$ (car toutes les dérivées d'ordre inférieur à k s'annullent). D'où $g \circ h(C_k \cap V) = f(C_k \cap V)$ est de mesure nulle. Comme $C_k \setminus C_{k+1}$ est couvert

par un nombre au plus dénombrable de tels ensembles V_i il suit que $f(C_k \setminus C_{k+1})$ est de mesure nulle.

#3 : Soit $I^n \subset U$ un cube de dimension n et de côté δ. Si k est assez grand c-à-d $k > \frac{n}{p} - 1$, on va prouver que $f(C_k \cap I^n)$ est de mesure nulle (2). Comme C_k peut être couvert par un nombre au plus dénombrable de tels cubes, on aura que $f(C_k)$ est de mesure nulle. Reste donc à vérifier (2).

Soit $x \in C_k$. Par compacité de I^n, on peut construire un développement de Taylor de f en x : $f(x+h) = f(x) + \frac{h^{k+1}}{(k+1)!} f^{(k+1)}(x) + \mathcal{O}(h^{k+2})$.

D'où, $f(x+h) = f(x) + R(x,h)$ (3) où $\|R(x,h)\| \leq C \|h\|^{k+1}$ (4)

$x \in C_k \cap I^n$, $x+h \in I^n$.

Subdivisons I^n en n^m cubes de côtés δ/n (voir fig. 2).

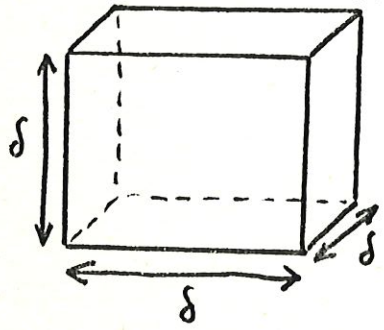
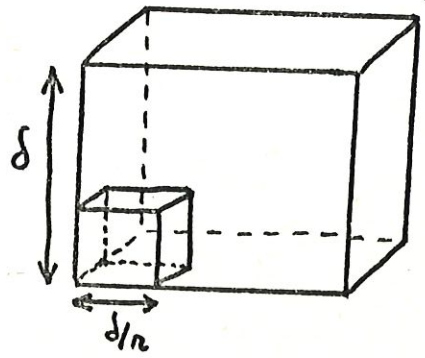


fig. 2



Soit I_1 un cube de cette subdivision contenant $x \in C_k$.

Alors, tout point de I_1 s'exprime comme $x+h$ où $\|h\| \leq \sqrt{n}\,(\delta/n)$ (5)

En effet, voyons cela pour les cas où $n \leq 3$.

$\underline{n=1}$: $x \in I_1$ et $x+h \in I_1$ donc, $\|h\| \leq \delta/n$

$\underline{n=2}$: Traitons le cas le plus défavorable (fig. 3)

Alors, $\|h\|^2 = (\delta/r)^2 + (\delta/r)^2$ et donc $\|h\| = \sqrt{2}\,\frac{\delta}{r}$.

En général, $\|h\| \leq \sqrt{2} \cdot \frac{\delta}{r}$.

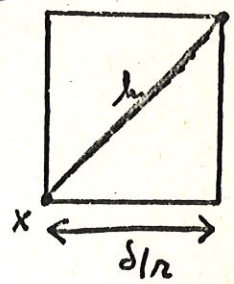


fig. 3

$\underline{n=3}$: Regardons de nouveau le cas le plus défavorable (fig. 4).

Alors, $\|h\|^2 = 2\left(\frac{\delta}{r}\right)^2 + \left(\frac{\delta}{r}\right)^2$ et par conséquent $\|h\| = \sqrt{3}\,\frac{\delta}{r}$. En général, $\|h\| \leq \sqrt{3} \cdot \frac{\delta}{r}$

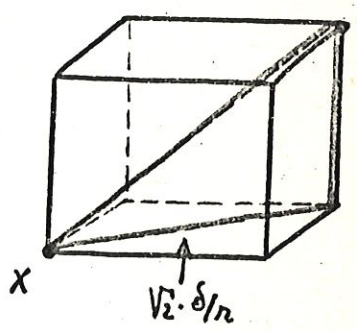


fig. 4

Le cas n général se traite de même façon ; on peut donc affirmer (5).

Par (3), (4) et (5), on a que la distance entre un point quelconque $f(x+h)$ de $f(I_1)$ et $f(x)$ est au plus $c\,\frac{(\sqrt{n}\,\delta)^{k+1}}{r^{k+1}}$. On peut alors conclure que $f(I_1)$ se trouve dans un cube centré sur $f(x)$ de côté a/r^{k+1} où $a = 2c(\sqrt{n}\,\delta)^{k+1}$.

D'où, $f(C_k \cap I^n)$ contient une union d'au plus r^n cubes ayant un volume $V \leq r^n (a/r^{k+1})^p = a^p\, r^{n-(k+1)p}$. Si $k+1 > n/p$ c-à-d $k > \frac{n}{p} - 1$, alors ce volume tend vers 0 si r augmente infiniment.

Donc, $f(C_k \cap I^n)$ aura une mesure nulle. Ceci terminera le pas #3 et achève la démonstration du théorème de Sard. ∎

Le théorème suivant est une conséquence importante du théorème de Sard.

II.1.3. Le théorème paramétrisé de Sard

Soient $V \subset \mathbb{R}^m$ et $U \subset \mathbb{R}^{n+1}$ des ouverts et $\Phi : V \times U \longrightarrow \mathbb{R}^p$ une application régulière. Si $0 \in \mathbb{R}^p$ est valeur régulière de Φ, alors pour presque tout $a \in V$, 0 est une valeur régulière de $\Phi_a(\cdot) \equiv \Phi(a, \cdot)$.

Nous nous intéressons en particulier au cas où $n = p$. Dans la suite, la notation $\langle\langle \cdot \rangle\rangle$ désigne "l'espace engendré par".

Notre hypothèse assure que 0 est valeur régulière de Φ. Comme en plus $\langle\langle D_{(a,x)} \Phi \rangle\rangle = \langle\langle D_a \Phi \rangle\rangle + \langle\langle D_x \Phi \rangle\rangle$, il sera suffisant de dire que $\langle\langle D_a \Phi \rangle\rangle = \mathbb{R}^p$ où $D_a \Phi$ représente la matrice des dérivées partielles de Φ par rapport à a (matrice de la forme $n \times p$).

Avant de démontrer le thm II.1.3., rappelons un résultat fondamental d'analyse (voir par exemple [7] et [13])

Théorème II.1.4. (Théorème des fonctions implicites)

Soit $f : E \subset \mathbb{R}^p \times \mathbb{R}^m \longrightarrow \mathbb{R}^p$ une application C^1. Supposons qu'en $(w^*, u^*) \in E$ $f(w^*, u^*) = 0_{\mathbb{R}^p}$ et $D_w f(w^*, u^*)$ est non-singulière.

Alors, il existe une boule ouverte $U \subset \mathbb{R}^m$ telle que $U = \{u \in \mathbb{R}^m \mid |u^* - u| < \delta\}$ et une application $g : U \longrightarrow \mathbb{R}^p$ de classe C^1 telle que

(i) $g(u^*) = w^*$ et (ii) $f(g(u), u) = 0_{\mathbb{R}^p} \quad \forall u \in U$

Le théorème peut se généraliser : Supposons que $f(w^*, u^*)$ quelconque et $D_w f(w^*, u^*)$ non-singulière. Alors, il existe U voisinage de u^*, V voisinage de $f(w^*, u^*)$ et une application unique $g : V \times U \to \mathbb{R}^p$ telle que pour tout $(v, u) \in V \times U$ $\quad f(g(v,u), u) = v$.

Preuve du thm II.1.3.

Par le thm II.1.4., on a que $M \equiv \Phi^{-1}(0)$ est surface régulière de $\mathbb{R}^n \times \mathbb{R}^{n+1}$. En effet, il faut que tout élément de M admet un voisinage $W \cap M$ difféomorphe à un sous-espace ouvert de $\mathbb{R}^{2n+1}$. Comme 0 est valeur régulière de Φ, on peut, par non-singularité de $D_a \Phi$, appliquer le thm II.1.4. Il suit qu'il existe un voisinage $W \cap M = \{(a,x) = (g(x), x) \in \mathbb{R}^{2n+1} \mid |x - x^*| < \delta, \delta > 0\}$ tel que pour tout $(a,x) \in W \cap M$ $\Phi(a,x) = 0$ où (a^*, x^*) sont tels que $\Phi(a^*, x^*) = 0$ et $g(a^*) = x^*$. Le voisinage est bien contenu dans $\Phi^{-1}(0)$, contient (a^*, x^*) et est difféomorphe à lui-même c-à-d à un ouvert de $\mathbb{R}^{2n+1}$.

Soit la projection $\pi : M \to \mathbb{R}^n$ telle que $\pi(a,x) = a$. Le thm I.1.2 entraîne que presque tout $a \in \mathbb{R}^n$ est valeur régulière de π. Pour tout $(a,x) \in \pi^{-1}(a) \subset \Phi^{-1}(0) = M$, on a que $\operatorname{rg}(D\pi(a,x)) = n$. On engendre par conséquent tout point de $\mathbb{R}^n$ à partir de (a,x) se trouvant sur

M c-à-d $\Phi(a,x)=0$. Donc, pour tout $b\in\mathbb{R}^n$ $\exists y\in\mathbb{R}^{n+1}$ tel que (b,y) est tangent à M ce qui implique que $D\Phi(a,x)\cdot(b,y)=0$.

$$\text{Or, } \begin{pmatrix} \frac{\delta\Phi_1}{\delta a_1} & \dots & \frac{\delta\Phi_1}{\delta a_m} & \frac{\delta\Phi_1}{\delta x_1} & \dots & \frac{\delta\Phi_1}{\delta x_{n+1}} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\delta\Phi_p}{\delta a_1} & \dots & \frac{\delta\Phi_p}{\delta a_m} & \frac{\delta\Phi_p}{\delta x_1} & & \frac{\delta\Phi_p}{\delta x_{n+1}} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_m \\ y_1 \\ \vdots \\ y_{n+1} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\delta\Phi_1}{\delta a_1} & \dots & \frac{\delta\Phi_1}{\delta a_m} \\ \vdots & & \vdots \\ \frac{\delta\Phi_p}{\delta a_1} & \dots & \frac{\delta\Phi_p}{\delta a_m} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} + \begin{pmatrix} \frac{\delta\Phi_1}{\delta x_1} & \dots & \frac{\delta\Phi_1}{\delta x_{n+1}} \\ \vdots & & \vdots \\ \frac{\delta\Phi_p}{\delta x_1} & & \frac{\delta\Phi_p}{\delta x_{n+1}} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_{n+1} \end{pmatrix}$$

$$= D_a\Phi(a,x)\cdot b + D_x\Phi(a,x)\cdot y = 0$$

Ceci entraîne une dépendance linéaire entre les matrices. Par la régularité de $0\in\mathbb{R}^p$ ($\operatorname{rg}(D_a\Phi(a,x)) = p$) on peut déduire que $\operatorname{rg} D_a\Phi(a,x) = \operatorname{rg} D_x\Phi(a,x) = \operatorname{rg} D\Phi(a,x) = p = n$. ∎

II.2. Le théorème de transversalité

Nous allons généraliser dans la suite les résultats obtenus au paragraphe II.1. Le théorème de transversalité représente ainsi une généralisation du thm paramétrisé de Sard. Il sera d'abord utile de se rappeler quelques notions préliminaires de transversalité.

dans $\mathbb{R}^3$ sont transversaux.

Nous généralisons cette notion aux applications et aux sous-variétés.

L'application $f_* = T_x f : T_x X \longrightarrow T_{f(x)} Y$ représente l'application linéaire tangente de $f : X \to Y$ où $T_x X$ est l'espace tangent de X en x. Par notation, prenons $Df(x) = T_x f = f_*$.

3. Soient A, B, C trois variétés différentielles et deux applications de classe C^∞ $f : B \to A$ et $g : C \to A$.

On dira que les applications f et g sont <u>transversales aux points</u> <u>$(b,c) \in B \times C$</u> si soit $f(b) \neq f(c)$

soit $f(b) = f(c) = a \in A$

et les images des espaces tangents à B et C sont transversales à A en a c-à-d $f_* T_b B + g_* T_c C = T_a A$.

Notons alors $f \pitchfork g$ en (b,c).

Deux applications sont dites <u>transversales</u> si elles sont transversales en tout couple de points (b,c).

Deux courbes dans $\mathbb{R}^3$ par exemple sont transversales si et seulement si elles ne se coupent pas.

4. Soit $f : B \to A$ une application de classe C^∞ et une sous-variété $C \subset A$. On dira que <u>f est transversale à la sous-variété C</u> si elle est transversale au plongement $i : C \to A$; on le note $f \pitchfork C$.

On aura donc, puisque le cas $f(b) \neq i(c) = c \in C \subset A$ pour tout (b,c) est impossible, que $f_* T_b B + i_* T_c C = T_c A$

c-à-d $\quad f_* T_b B + T_c C = T_c A \quad$ avec $f(b) = i(c) = c$.

Considérons $f : \mathbb{R}^n \to \mathbb{R}^p$, $n \geq p$, une application régulière. L'ensemble des valeurs régulières de f est dense dans $\mathbb{R}^p$ (thm de Sard). On aura donc que pour presque tout $c \in \mathbb{R}^p$ $\operatorname{rg}(Df(b)) = p$ où $f(b) = c$. Ceci implique donc que f est transversale à $\mathbb{R}^p$.

Si $f : X \to Y$ est de classe C^∞, alors $f_* T_x X \neq T_{f(x)} f(X)$. Ainsi, une courbe du plan peut ne pas être transversale à une courbe donnée même quand son image est normale à la courbe donnée. Prenons par exemple $f : \mathbb{R} \to \mathbb{R}^2$ tel que $f(x) = (x^3, x^3)$. On a que $f_*(x) = (3x^2, 3x^2)$. On a que $f_* T_0 X = \{(0,0)\}$ tandis que $T_{(0,0)} f(x) = \{(x,x) \in \mathbb{R}^2\}$.

On étudie dans la suite une série de propositions assez techniques tirées de [6] qui vont être à la base du thm de transversalité (thm II.2.6.).

Proposition II.2.2.

Soient X et Y deux variétés différentielles et $W \subset Y$ une sous-variété. Supposons que $\dim W + \dim X < \dim Y$ c-à-d $\dim X < \operatorname{codim} W$. Soit $f : X \to Y$ de classe C^∞ et $f \pitchfork W$. Alors, $f(X) \cap W = \phi$.

Preuve : Le thm explicite l'importance des dimensions relatives de X, Y et W dans le problème de transversalité. Supposons par l'absurde que $f(x) \in W$ pour un certain $x \in X$. Puisque $f \pitchfork W$, $\langle\langle Df(x) \rangle\rangle + T_y W = T_y Y$ pour tout $y = f(x) \in W$. Comme $\dim \left(T_{f(x)} W + f_* T_x X \right) \leq \dim T_{f(x)} W + \dim T_x X = \dim W + \dim X < \dim Y = \dim T_{f(x)} Y$, il sera impossible que $T_{f(x)} W + f_* T_x X = T_y Y$ (on contredit l'hypothèse $f \pitchfork W$). ∎

Etant donné une fonction $f : U \subset X \to Y$ de classe C^1. On dit que f est une <u>submersion de U</u> ssi $\forall x \in U \quad f_* = T_x f : T_x X \to T_{f(x)} Y$ est surjective

Par le thm de Sard, une application régulière $f : \mathbb{R}^n \to \mathbb{R}^p$, $n \geq p$ est une submersion de $\mathbb{R}^n$. Pour cela il suffit de remarquer que la matrice jacobienne de f est de rang plein en l'image inverse de presque tout point de $\mathbb{R}^n$.

Proposition II.2.3.

Soient X et Y deux variétés différentielles et $W \subset Y$ une sous-variété de Y. Supposons en plus f de classe C^∞, $p \in X$, $f(p) \in W$ et qu'il y a un voisinage U de $f(p)$ dans Y et une submersion $\phi : U \to \mathbb{R}^k$ ($k = \operatorname{codim} W$) telle que $W \cap U = \phi^{-1}(0)$.

Alors, $f \pitchfork W$ en $p \iff \phi \circ f$ est une submersion en p.

Preuve: Un tel voisinage U existe toujours. En effet, $f(p) \in W$ et W est une sous-variété implique par définition qu'il existe une carte $c = (U, \Psi, m)$ de W en $f(p)$ telle que $W \cap U = \Psi^{-1}(V)$ (où $\Psi : U \subset W \to V \subset \mathbb{R}^m$).

Comme $W \cap U = \phi^{-1}(0)$, prenons $V = \{(0, \dots, 0, x_{k+1}, \dots, x_m) \in \mathbb{R}^m = \mathbb{R}^k \times \mathbb{R}^{m-k}\}$.

Il suffit de prendre $\phi = \pi \circ \Psi$ où $\pi : \mathbb{R}^m = \mathbb{R}^k \times \mathbb{R}^{m-k} \to \mathbb{R}^k$ est la projection canonique. Considérons le diagramme de la fig. 6 :

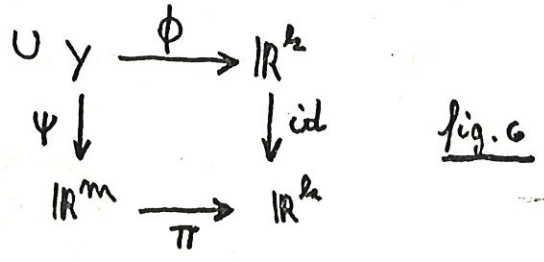
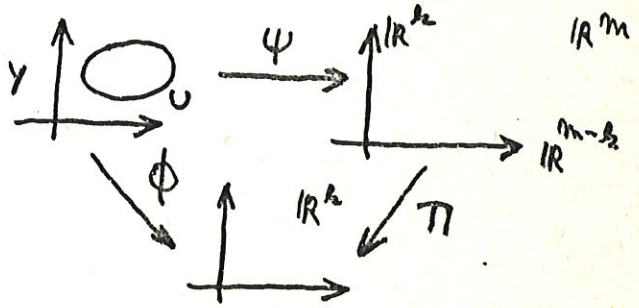


fig. 6



il est évident que π n'est rien d'autre que l'expression locale de ϕ par rapport aux cartes $c = (U, \Psi, m)$ et $c' = (\mathbb{R}^k, id, k)$.

On aura pour l'application linéaire tangente le diagramme de fig. 7.

$$\begin{array}{ccc} T_{f(x)}Y & \xrightarrow{\phi_*} & T\mathbb{R}^k \\ \downarrow \theta_c & & \downarrow \theta_i \\ \mathbb{R}^m & \xrightarrow[\theta_c^{-1}\circ\Pi\circ\theta_i]{} & \mathbb{R}^k \end{array}$$

fig. 7

On, comme pour tout $p \in W$ $\psi(p) = (0, \dots, 0, x_{k+1}, \dots, x_m)$ par définition de V, il est évident que pour $v \in T_{f(x)}W$: $\theta_c(V) = (0, \dots, 0, v_{k+1}, \dots, v_m)$ de sorte que $\phi_*(V) = 0$ si $v \in T_{f(p)}W$ (on ne considère que les k premières composantes). Ceci implique que $T_{f(p)}W \subset \ker[T_{f(p)}\phi]$ (i). Mais, $\dim T_{f(p)}W = \dim W = m-k$ par hypothèse (ii) et comme ϕ est une submersion en $f(p)$ on a que $\operatorname{rg}(T_{f(p)}\phi) = k$. Par conséquent, on peut déduire que le $\operatorname{rg}[\ker(T_{f(p)}\phi)] = m-k$ (iii). D'où, par (i), (ii) et (iii) $\ker(T_{f(p)}\phi) = T_{f(p)}W$. Le fait que $f \pitchfork W$ s'exprime par $T_{f(p)}Y = T_{f(p)}W + f_* T_p X \iff T_{f(p)}Y = \ker(T_{f(p)}\phi) + f_* T_p X$ c-à-d $T_{f(p)}Y = \ker(\phi_*)_{f(p)} + f_* T_p X$ (0). Puisque $T_{f(p)}\phi$ est surjectif (de rg k), $T_p(\phi \circ f)$ le sera $\iff$ (0) est vérifié.

En effet, $T_p X \xrightarrow{f_*} T_{f(p)}Y \xrightarrow{\phi_*} T_{(\phi\circ f)(p)}\mathbb{R}^k$. Le fait que ϕ_* est surjectif et (0) entraînent que $T_p(\phi \circ f)$ est surjectif. Si $T_{f(p)}Y \neq \ker T_{f(p)}\phi + f_* T_p X$, il existera un $V \neq 0$, $V \in T_{(\phi\circ f)(p)}\mathbb{R}^k$ tel que $V = \phi_*(W)$ pour un certain $w \in T_{f(p)}Y - f_* T_p X$ de sorte que $T_p(\phi \circ f)$ ne serait pas surjectif.

On conclut que $\phi \circ f$ est une submersion $\iff f \pitchfork W$ en p. ∎

Proposition II.2.4.

Soient X et Y des variétés différentiables et W une sous-variété de Y. Supposons également f de classe C^∞ et $f \pitchfork W$.
Alors, $f^{-1}(W)$ est une sous-variété de X.

Preuve : Soient $p \in W$, $c = (U, \psi, n)$ une carte de Y en p et ϕ de U vers $\mathbb{R}^k$ une submersion telle que $W \cap U = \phi^{-1}(0)$ ainsi que V un voisinage de p tel que $f(V) \subset U$. Par la proposition II.2.3, $\phi \circ f$ est une submersion sur V c-à-d $\operatorname{rg}(T(\phi \circ f)(p)) = k$. Puisque $f(V) \subset U$, on a que $V \subset f^{-1}(U)$ et on peut se restreindre à V.
$W \cap U = \phi^{-1}(0)$ entraîne $f^{-1}(W \cap U) = f^{-1}(\phi^{-1}(0))$; d'où, $f^{-1}(W) \cap f^{-1}(V) = (\phi \circ f)^{-1}(0)$ ce qui donne (se restreindre à V) alors $f^{-1}(W) \cap V = (\phi \circ (f/V))^{-1}(0)$. Or, l'image réciproque d'une sous-variété par une submersion est une sous-variété. D'où, la thèse. ∎

Proposition II.2.5.

Soient X, B et Y des variétés différentielles et W une sous-variété de Y. Soit $j : B \longrightarrow \{ f \mid f : X \to Y \text{ de classe } C^\infty \}$ une application pas nécessairement continue et $\Phi : X \times B \longrightarrow Y$ tel que $\Phi(x, b) = j(b)(x)$
Supposons Φ de classe C^∞ et $\Phi \pitchfork W$.

Alors, l'ensemble $\{b \in B \mid j(b) \pitchfork W\}$ est dense dans B.

Preuve: Soit $W_\phi = \phi^{-1}(W)$. Par II.2.4 et puisque $\phi \pitchfork W$, W_ϕ est une sous-variété de $X \times B$. Soit π la restriction à W_ϕ de la projection canonique de $X \times B \to B$ c-à-d $\pi : \phi^{-1}(W) \subset X \times B \to B$ tel que $\pi(x,b) = b$. Notons d'abord que si $b \notin \operatorname{Im} \phi$, alors $\phi(x,b) \cap W = j(b)(x) \cap W = 0$ et $j(b) \pitchfork W$ (car $\phi \pitchfork W$, $b \notin \operatorname{Im} \phi$ et en appliquant la proposition II.2.2.). Si $\dim W_\phi < \dim B$, $\pi(W_\phi)$ est de mesure nulle dans B et il y a une partie dense de B, à savoir $B - \operatorname{Im} \pi$, telle que $j(b) \pitchfork W$ si b lui appartient. Dans ce cas, la proposition est donc vérifiée et on peut alors supposer que $\dim W_\phi \geq \dim B$. Montrons maintenant que, si b est une valeur régulière pour π, alors $j(b) \pitchfork W$. Si cette assertion est vérifiée, la proposition découle du thm de Sard (I.1.2.) appliqué à π à savoir que l'ensemble des valeurs régulières de π est dense dans B.

Supposons donc que b soit une valeur régulière de π et soit $x \in X$. Si $(x,b) \notin W_\phi$, alors $j(b)(x) \notin W$ et $j(b) \pitchfork W$ ($\Rightarrow \blacksquare$). On peut donc supposer $(x,b) \in W_\phi$. Comme b est une valeur régulière de π (ce qui implique que π est transversale à B) et que $\dim W_\phi \geq \dim B$, on a

que $\phi(x,b) \in W$ c-à-d $j(b)(x) \in W$ et en plus que $T_{(x,b)}{}^{X\times B} =$ $T_{(x,b)}{}^{W\phi} + T_{(x,b)}{}^{X\times\{b\}}$. Le fait que π est une submersion implique $T_{(x,b)}\{x\}\times B \subset T_{(x,b)}W\phi$ (par surjectivité de $T_{\pi(x,b)}{}^{W\phi}$). Appliquons ϕ_* aux deux côtés de cette égalité et nous obtenons que $\phi_*\, T_{(x,b)}{}^{X\times B} = T_{j(b)(x)}\, W + (j(b))_*\, T_x X$ (a). En effet, ceci vient de $\phi_*\, T_{(x,b)}\, \phi^{-1}(W) = T_{\phi(x,b)}{}^{W} = T_{j(b)(x)}W$ et $\phi_*\, T_{(x,b)}{}^{X\times\{b\}} =$ $D_x\phi\, T_x X \times D_b\phi\, T_b\{b\} = D_x j(b)\, T_x X$. On a supposé que $\phi \pitchfork W$ de sorte que $T_{j(b)(x)}Y = T_{\phi(x,b)}Y = T_{\phi(x,b)}W + (T_{(x,b)}\phi)(T_{(x,b)}{}^{X\times B})$ (b)

En combinant les égalités (a) et (b), on obtient que

$T_{j(b)(x)}Y = T_{j(b)(x)}\, W + (T_x j(b))\, T_x X$ ce qui n'est rien d'autre que $j(b) \pitchfork W$.

Remarque importante :

Soient $G : X\times B \to Y$ une famille d'applications de $X \to Y$ paramétrisée par B telle que $G_b(x) \equiv G(x,b)$ et $j : B \to \{f : X \to Y$ de classe $C^\infty\}$ donnée par $j(b) \equiv G_b$.

Alors, $\phi = G$. Supposons que $G \pitchfork W$. Dans ce cas, l'ensemble $\{b \in B \mid G_b \pitchfork W\}$ est dense dans B.

Cette remarque est un résultat fondamental de la transversalité. Si une famille paramétrisée d'applications est transversale à une sous-variété donnée, alors les applications individuelles seront également transversales à cette variété, au moins pour une partie dense de l'espace des paramètres.

Ce résultat très important est formulé dans le

II.2.6. Le théorème de transversalité

Soient X, Y et B des variétés différentielles de dimension $n+1$, m et p. Soit en plus $W \subset Y$ une sous-variété et une application régulière (ici C^∞) $\phi : B \times X \to Y : (a, \lambda, x) \to \Phi(a, \lambda, x)$.
Si Φ est transversale à W, alors pour presque tout $a \in B$
$\Phi_a(\cdot) = \Phi(\lambda, x) : X \to Y$ est transversale à W.

Il représente une généralisation du théorème paramétrisé de Sard décrit au paragraphe II. 1.

II.3. Caractérisation de $\Phi_a^{-1}(0)$

Travaillons dans la suite avec le cas $n = p$. On est maintenant capable de caractériser la courbe nulle de $\Phi_a(\lambda, x)$. Ceci sera fait sous forme d'un corollaire des théorèmes II.1.3 et II.2.6.

Corollaire II.3.1.

Reprenons les mêmes hypothèses qu'aux théorèmes II.1.3 et II.2.6. Alors, toute composante de $\Phi_a^{-1}(0)$ (resp. $\Phi_a^{-1}(W)$) est une courbe régulière pour presque tout $a \in V$ (resp. $a \in X$).

On va établir la preuve dans le cas de $\Phi_a^{-1}(0)$. L'autre situation se déduit de façon analogue en se basant sur le théorème de transversalité et la généralisation du théorème des fonctions implicites.

Preuve : Reste à voir qu'il existe pour tout $x \in X$ un voisinage $\mathcal{O} \cap \Phi_a^{-1}(0)$ difféomorphe à un intervalle ouvert de $\mathbb{R}$.

Or, $\Phi : V \subset \mathbb{R}^n \times U \subset \mathbb{R}^{n+1} \to \mathbb{R}^p$ est régulière (de classe C^1, C^2 ou (∞)). Comme 0 est valeur régulière de Φ (thm de Sard) on déduit par le

thm paramétrisé de Sard (II.1.3) que 0 est valeur régulière de Φ_a c-à-d que pour tout $(\lambda,x) \in \Phi_a^{-1}(0)$ $\operatorname{rg}(D\Phi_a(\lambda,x)) = n$ où $D\Phi_a(\lambda,x)$ est le jacobien de $\Phi_a(\lambda,x)$ (matrice $n \times n+1$). Prenons un point (λ^*,x^*) quelconque tel que $\Phi_a(\lambda^*,x^*)=0$; comme le jacobien est de rang plein on peut appliquer le théorème des fonctions implicites. Il existe par conséquent un ouvert $\Lambda^* = \{\lambda \in \mathbb{R} \mid |\lambda-\lambda^*| < \rho\}$ et $g : \Lambda^* \to \mathbb{R}^n$ tels que $g(\lambda^*) = x^*$ et $\forall \lambda \in \Lambda^*$ $g(\lambda) = x$, $\Phi_a(\lambda,x) = \Phi(a,\lambda,x) = 0$.
Ceci peut être fait pour tout $(\lambda^*,x^*) \in \Phi_a^{-1}(0)$ par la régularité de 0.
Notre raisonnement restera donc valable pour toute composante de $\Phi_a^{-1}(0)$.
Or, $g : \Lambda^* \to g(\Lambda^*) \subset \mathbb{R}^n$ tel que $\Phi_a(\lambda, g(\lambda)) = 0$.
On peut donc décrire toute composante de $\Phi_a^{-1}(0)$ par ces voisinages $(\Lambda, g(\Lambda))$ où $\Lambda \equiv \cup\{\Lambda^* \mid (\lambda^*,x^*)$ appartient à cette composante de $\Phi_a^{-1}(0)\}$. On établit ainsi un difféomorphisme $h^{\leftarrow}$ entre $\Phi_a^{-1}(0)$ et Λ (ouvert de $\mathbb{R}$) :

$$h : \lambda \in \Lambda \to (\lambda, g(\lambda))$$
$$h^{\leftarrow} : (\lambda, g(\lambda)) \to \lambda \in \Lambda$$
où $(\lambda, g(\lambda)) \in \Phi_a^{-1}(0)$

Ainsi, toute composante de $\Phi_a^{-1}(0)$ est une courbe régulière. ∎

D'où, on peut conclure que si on choisit un point par hasard dans V, on a une probabilité égale à un que toute composante de $\Phi_a^{-1}(0)$

Soit une courbe régulière.

En pratique, en connaissant un point p de $\Phi_a^{-1}(0)$ (souvent $(0,a)$) nous désignons par Γ_a la composante de $\Phi_a^{-1}(0)$ qui le contient.

Plus généralement, nous sommes capables de distinguer une composante particulière de $\Phi_a^{-1}(0)$ qu'on notera Γ_a (voir fig. 8). Nous construisons dans la suite un algorithme qui suivra cette courbe Γ_a.

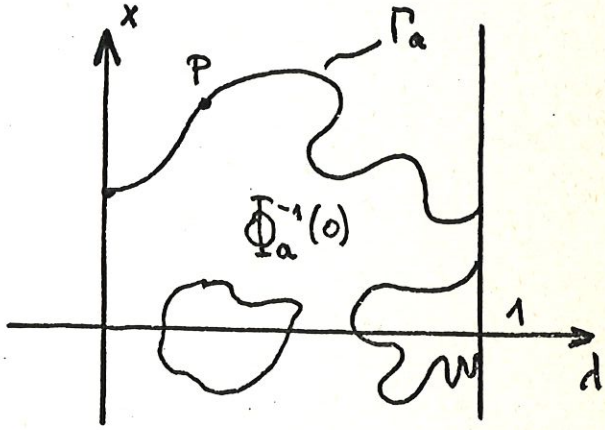


fig. 8

Soit $f : \bar{B}^n \to \bar{B}^n$ régulière. Pour $\bar{B}^n$, on peut aussi bien prendre une boule $k\cdot\bar{B}^n$ quelconque. Au moyen de la fonction homotopique, on peut soit traiter le problème de points fixes (i), soit celui de racines (ii)

Prenons $g_a(x) = x - a$.

Dans le cas (i) $\Phi_a(\lambda, x) \equiv \Phi(a, \lambda, x) = (1-\lambda)(x-a) + \lambda(x - f(x))$ (6)

(ii) $\Phi_a(\lambda, x) \equiv \Phi(a, \lambda, x) = (1-\lambda)(x-a) + \lambda f(x)$ (7)

Nous traitons dans la suite le cas (i) c-à-d la recherche de points fixes de f.

Soit Γ_a la composante de $\Phi_a^{-1}(0) \cap [0,1] \times \bar{B}^n$ contenant $(0,a)$.

Il suit maintenant le <u>théorème fondamental</u> de ce chapitre qui donne une bonne caractérisation de la courbe Γ_a.

Théorème II.3.2.

Considérons l'application $\Phi : \overline{B}^n \times (0,1) \times \overline{B}^n \longrightarrow \mathbb{R}^n$ définie par (6). Alors, (a) $0 \in \mathbb{R}^n$ est valeur régulière de Φ.

(b) Pour presque tout $a \in \overline{B}^n$, Γ_a est une courbe régulière dans $(0,1) \times \overline{B}^n$ liant $(0,a)$ à un point fixe ou un ensemble de points fixes en $\lambda = 1$.

En pratique Γ_a représente souvent un arc régulier menant à un point fixe de f. Il y a pourtant des possibilités exigeant que Γ_a ne rencontre pas un point fixe, mais converge vers un ensemble de points fixes (voir fig. 9). D'où, $\Phi_a^{-1}(0)$ est régulière si on considère sa restriction à $(0,1) \times \overline{B}^n$ ou à $[0,1) \times \overline{B}^n$, mais pas nécessairement à $[0,1] \times \overline{B}^n$.

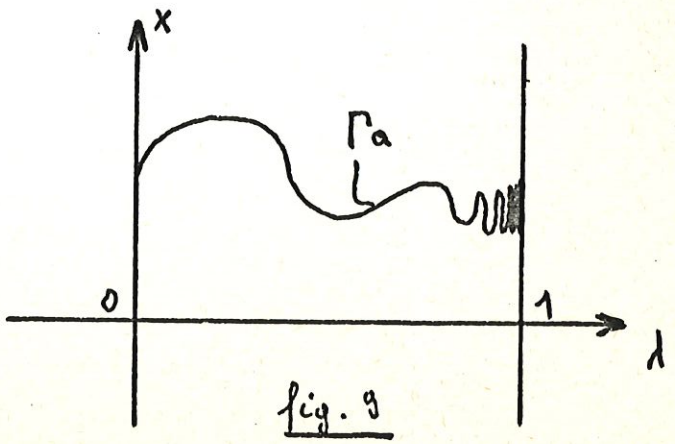


fig. 9

Avant de passer à la preuve du théorème II.3.2. rappelons un résultat élémentaire de la géométrie différentielle (voir aussi [5])

Lemme II.3.3.

Toute variété régulière de dimension un est difféomorphe soit à un cercle soit à un intervalle ouvert.

Preuve du thm II.3.2.

(a) Soit $(\bar{a}, \bar{\lambda}, \bar{x}) \in \bar{B}^n \times (0,1) \times \bar{B}^n$ tel que $\Phi(\bar{a}, \bar{\lambda}, \bar{x}) = 0$. Alors, $D_a \Phi(\bar{a}, \bar{\lambda}, \bar{x}) = -(1-\lambda) I$ où $I \equiv$ matrice identité $n \times n$. Donc, pour $\lambda \neq 1$ $\langle\langle D_a \Phi(\bar{a}, \bar{\lambda}, \bar{x}) \rangle\rangle = \mathbb{R}^n$. Or $\langle\langle D_a \Phi(\bar{a}, \bar{\lambda}, \bar{x}) \rangle\rangle \subset \langle\langle D\Phi(\bar{a}, \bar{\lambda}, \bar{x}) \rangle\rangle$. D'où, pour tout $(\bar{a}, \bar{\lambda}, \bar{x}) \in \Phi^{-1}(0)$ $rg\,(D\Phi(\bar{a}, \bar{\lambda}, \bar{x})) = n$ c-à-d $0 \in \mathbb{R}^n$ est valeur régulière de Φ.

(b) Le thm II.3 implique que 0 est valeur régulière de Φ_a. Par le corollaire II.3.1. toute composante de $\Phi_a^{-1}(0)$ est une courbe régulière dans $(0,1) \times \bar{B}^n$. Par le lemme II.3.3. il existe un difféomorphisme à un cercle ou à un intervalle ouvert. Considérons l'application Φ_a de domaine $(-\alpha, 1) \times \mathbb{R}^n$. Or $\Phi_a(0,a) = 0$ et $D_x \Phi_a(0,a) = I$. Soit $|\lambda| \ll 1$ dans un voisinage de $(0,a)$;

l'application du thm des fonctions implicites en ce voisinage implique l'existence de l'application $x : (0,1) \longrightarrow \overline{B}^n$ telle que $\Phi_a(\lambda, x(\lambda)) = 0$; $x(0) = a$. Γ_a sera par conséquent difféomorphe à un intervalle ouvert (ici $(0,1)$) et non à un cercle.

Reste à voir que Γ_a n'a pas de points limites à la surface latérale $(0,1) \times S^{n-1}$. En effet, sinon une telle limite (b, x) vérifierait $\Phi_a(\lambda, x) = 0$, $0 < \lambda < 1$ c-à-d $x = (1-\lambda)\, a + \lambda f(x)$. D'où, x se trouve sur le segment de droite entre a et $f(x)$. Comme $a \in B^n$, $f(x) \in \overline{B}^n$ (comme point limite), on a que $x \in B^n$. Or, $B^n \cap S^{n-1} = \phi$ et donc, Γ_a n'a pas de points limites dans $(0,1) \times S^{n-1}$.

Evidemment toute limite de Γ_a se trouve dans $\Phi_a^{-1}(0)$; il en est avec le seul point limite de Γ_a dans $\{0\} \times \overline{B}^n$ c-à-d $(0, a)$. De plus, on a vu que Γ_a était difféomorphe à un intervalle ouvert et $(0, a)$ se situe à une de ses fins. Par compacité de $[0,1] \times \overline{B}^n$, il doit y avoir un point limite supplémentaire (ou plusieurs) dans $\{1\} \times \overline{B}^n$ ($(0, a)$ étant limite unique en $\{0\} \times \overline{B}^n$ par le choix de $g_a(x)$). Si $(1, \bar{x})$ est un tel point, alors $\bar{x}$ est un point fixe de f. ∎

II.3.4. Remarques

1. Si $I - Df(x)$ est non-singulière en tout point fixe de f, alors Γ_a est une courbe régulière dans $[0,1] \times \bar{B}^n$ et donc de longueur finie.

En effet, $D\Phi_a(\lambda, x) = [a - f(x), I - \lambda Df(x)]$ et prolongeons Φ_a à $[0, 1+d] \times \mathbb{R}^n$, $0 < d < 1$. Soit $\bar{x}$ un point fixe trouvé en $\lambda = 1$. Donc, $D\Phi_a(1, \bar{x}) = [a - f(\bar{x}), I - Df(\bar{x})]$ sera de rang plein. Je peux ainsi établir (théorème des fonctions implicites) un difféomorphisme local entre un ouvert de $\mathbb{R}$ et Γ_a. Γ_a est par conséquent de longueur finie au voisinage de $(1, \bar{x})$ et passer par $(1, \bar{x})$. D'où, la courbe sera de longueur finie sur $[0,1] \times \bar{B}^n$.

2. Le thm II.3.2. ne reste pas valable pour la recherche de racines.

En effet, dans ce cas $\Phi_a(\lambda, x) = (1-\lambda)(x-a) + \lambda f(x) = 0$. Reprenons la preuve du thm II.3.2. Pour un point limite $(\bar{\lambda}, \bar{x})$; $0 < \lambda < 1$, $f(\bar{x}) \in \bar{B}^n$ et $a \in B^n$ on a $(1-\lambda)\bar{x} = (1-\lambda)a - \lambda f(\bar{x})$ c-à-d que $\bar{x} = \dfrac{(1-\lambda)a - \lambda f(\bar{x})}{(1-\lambda)}$. Donc $\bar{x}$ ne se trouve pas nécessairement sur le segment de droite entre a et $f(\bar{x})$. D'où, le point $\bar{x}$ n'est pas obligé de rester dans B^n. Il peut aller à l'∞ (si $\lambda \to 1$)

Ainsi l'algorithme pour la recherche de racines apporte des

modifications. Il permet par exemple de sortir de $\bar{B}^n$. Le cas sera étudier dans le chapitre IV où on traitera un problème particulier de recherche de racines.

3. La courbe Γ_a ne peut jamais se couper avec elle-même. En effet, dans ce cas il n'y a plus moyen d'établir un difféomorphisme local au voisinage du point d'intersection entre Γ_a et un ouvert de $\mathbb{R}$. Or ce difféomorphisme devrait exister puisque Γ_a est une courbe régulière (voir fig. 10).

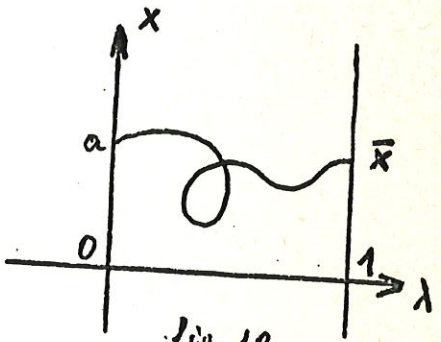


fig. 10

4. Le thm II.3.2. reste valable dans le cas où on généralise les boules (rayon arbitraire). Soit $\bar{B}^n_{k_2} = \{ x \in \mathbb{R}^n \mid |x| \leq k_2 \ , k_2 \in \mathbb{R} \}$ une boule de rayon k_2 quelconque. Ainsi pour des applications de $\bar{B}^n_{k_2}$ vers $\bar{B}^n_{k_2}$ $\Phi_a^{-1}(0)$ restera dans $\bar{B}^n_{k_2}$ et trouvera un point fixe de la fonction considérée.

II.4. Le théorème du point fixe de Brouwer

Du thm II.3.2. on peut déduire le théorème du point fixe dans le cas d'une fonction régulière (C^1, C^2 ou C^∞).

Lemme II.4.1.

Toute application régulière $f : \overline{B}^n \to \overline{B}^n$ admet un point fixe.

II.4.2. Théorème du point fixe de Brouwer

Toute application continue $f : \overline{B}^n \to \overline{B}^n$ admet un point fixe

Preuve : On se ramène au lemme II.4.1. en approximant f par une application régulière (un polynôme par exemple).

Soit donné $\varepsilon > 0$, par le thm d'approximation polynomiale de Weierstrass, il existe une fonction polynomiale $P_1 : \mathbb{R}^n \to \mathbb{R}^n$ telle $\|P_1(x) - f(x)\| < \varepsilon$ pour tout $x \in \overline{B}^n$. Pourtant, P_1 peut envoyer des points de $\overline{B}^n$ en dehors de $\overline{B}^n$. Pour éviter ceci, posons $P(x) = \frac{P_1(x)}{1+\varepsilon}$.

Donc, $P : \overline{B}^n \to \overline{B}^n$ tel que $\|P(x) - f(x)\| \leq \|P(x) - P_1(x)\| + \|P_1(x) - f(x)\| < \|P_1(x)\left(\frac{1}{1+\varepsilon} - \varepsilon\right)\| + \varepsilon \leq \|P_1(x)\| \left|\frac{\varepsilon}{1+\varepsilon}\right| + \varepsilon \leq \varepsilon'$ aussi petit qu'on veut.

Supposons par l'absurde que $f(x) \neq x$ pour tout $x \in \overline{B}^n$. Alors, la fonction continue $\|f(x) - x\|$ prendre un minimum $\mu > 0$ sur $\overline{B}^n$.

Soit $P(x)$ défini avant avec $\|f(x) - P(x)\| < \mu \quad \forall x \in \overline{B}^n$. Ceci implique que $P(x) \neq x$ pour tout $x \in \overline{B}^n$. D'où, $P(\cdot)$ est une application régulière de $\overline{B}^n$ vers $\overline{B}^n$ n'ayant pas de point fixe ce qui contredit le lemme II.4.1 ■

Le théorème suivant est une des versions multiples du théorème du point fixe de Brouwer où on traite cette fois des fonction de $\mathbb{R}^n$ dans $\mathbb{R}^n$.

Thm II.4.3.

Soit $f: \mathbb{R}^n \longrightarrow \mathbb{R}^n$ régulière. Supposons que $f(x) \cdot x \geq 0$ si $|x| = 1$. Alors, il existe $|x| \leq 1$ tel que $f(x) = 0$.

Preuve: Procédons par contradiction en donnant une preuve constructive analogue à celle du thm II.3.2.

Soit l'application $\Phi_a(\lambda, x) = (1-\lambda)(x-a) + \lambda f(x)$ où $|a| < 1$, $0 < \lambda < 1$.

Avec des arguments semblables à ceux du thm II.3.2. nous trouverons que pour presque tout $a \in \overline{B}^n$, il existe une courbe régulière Γ_a de $\Phi_a^{-1}(0)$. Reste à voir que Γ ne s'arrête pas en $|x| = 1$ et $\lambda < 1$, mais continue jusque $\lambda = 1$ c-à-d que Γ_a ne contienne pas de points

(λ, x) avec $|x|=1$ et $0<\lambda<1$. En effet, un tel point serait tel que $(1-\lambda)(x-a)+\lambda f(x)=0$. Or, $f(x).x \geq 0$; d'où il faut que $x(1-\lambda)(x-a)+\lambda f(x).x=0$ c-à-d que $x(x-a)\leq 0 \iff x^2 \leq ax$. Comme $\|x\|=1$, il faut $1 \leq \|a\|\,\|x\|$ c-à-d $1 \leq \|a\|$ ce qui est une contradiction avec l'hypothèse.

Comme Γ_a admet un point final en $(0, a)$, par compacité de $[0,1]$ et continuité de Φ, on trouvera au moins un autre point limite en $(1, \bar{x})$ où $\bar{x} \equiv$ zéro de f. ■

Le théorème implique qu'il y a une valeur intermédiaire de norme inférieure à 1, annulant f. (fig. 11)

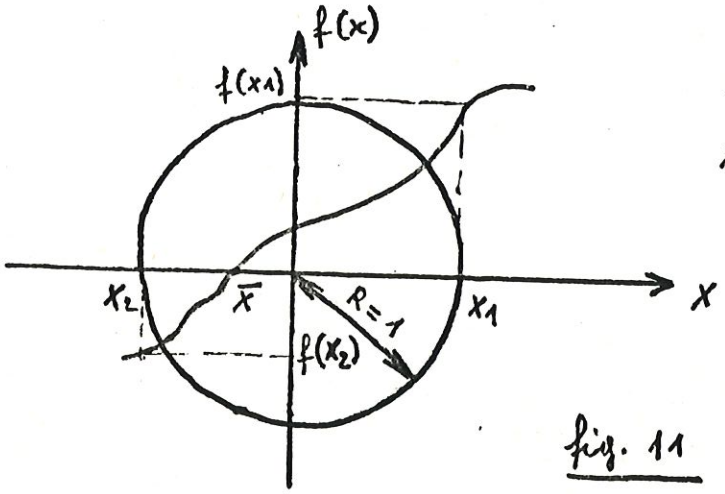


fig. 11

Il se pose maintenant le problème suivant : Comment faut-il suivre la courbe Γ_a en partant d'un point donné $(0, a)$ pour arriver à un point fixe (resp. racine) de f.

II.5. Suivre la courbe Γ_a

On montrera dans ce paragraphe que Γ_a est la solution d'une équation différentielle définie sur un ensemble ouvert. D'où, la courbe peut être suivie numériquement en utilisant un algorithme d'intégration d'une équation différentielle ($\in$D).

Soit $K(\lambda,x) \equiv \mathrm{Ker}\{D\Phi_a(\lambda,x)\} = \{ v \in \mathbb{R}^{n+1} \mid D\Phi_a(\lambda,x)\, v = 0 \}$ et $\Theta = \{(\lambda,x) \in [0,1] \times \bar{B}^n$ tel que le sous-espace $K(\lambda,x)$ est de dimension un$\}$. Prenons $\{e_1 \dots e_n\}$ une base vectorielle de $\mathbb{R}^n$, $(\lambda,x) \in \Theta$ et $H(\lambda,x)$ sous-espace de $\mathbb{R}^{n+1}$ de dimension n perpendiculaire à $K(\lambda,x)$. Comme (e_i) est une base de $\mathbb{R}^n$, $1 \le i \le n$, considérons $\tilde{e}_i$ base de $H(\lambda,x)$, $1 \le i \le n$, $\tilde{e}_i \in \mathbb{R}^{n+1}$: il peut y avoir une combinaison linéaire entre $\tilde{e}_i$ et e_i. Prenons donc les $\tilde{e}_i$ de façon à ce que $D\Phi_a(\lambda,x)\,\tilde{e}_i = e_i$ ce qui implique une dépendance linéaire entre $\tilde{e}_i$ et e_i. Remarquons l'unicité des $\tilde{e}_i$ par le fait que le jacobien $D\Phi_a(\lambda,x)$ est de rang plein.

Pour un $v \in \mathbb{R}^{n+1}$, soit $\mu(v) = \mu(v,\lambda,x)$ par définition égal au déterminant de l'ensemble ordonné $(v, \tilde{e}_1, \dots, \tilde{e}_n)$.

Si $\nu \in K(\delta,x)$, $\nu \neq 0$, il suit par perpendicularité de H et K que $\{\nu, \tilde{e_1}, \dots, \tilde{e_n}\}$ constitue une base de $\mathbb{R}^{n+1}$ c-à-d que $\mu(\nu) \neq 0$.

Soit un a fixé, par les thms II 1.2. et II. 1.1. 0 est valeur régulière de $\Phi_a(\delta,x)$ c-à-d que pour tout $(\delta,x) \in \Phi_a^{-1}(0)$ $\operatorname{rg}(D\Phi_a(\delta,x)) = n$.

Les éléments $\nu \in K(\delta,x)$ seront donc linéairement dépendents c-à-d $K(\delta,x)$ est de dimension un pour tout $(\delta,x) \in \Phi_a^{-1}(0)$. Ceci implique que $\mathcal{O}$ est un voisinage de $\Phi_a^{-1}(0)$ et donc aussi de Γ_a.

Il y a deux champs de vecteurs naturels : les équations différentiell sur $\mathcal{O}$. Soit $y \equiv (\delta,x)$, $\nu^+(y)$ le vecteur unité de $K(y)$ pour lequel $\mu(\nu^+(y)) > 0$ et $\nu^-(y)$ l'autre vecteur unité $(-\nu^+(y))$.

Si y est un point de Γ_a (on sait que $y \in \Phi_a^{-1}(0)$) alors $\nu^+(y)$ et $\nu^-(y)$ sont tous les deux tangents à Γ_a en y (en effet, on a que $D\Phi_a(y)\,\nu(y) = 0$). Ecrivons $\nu(y)$ pour ν^+ ou bien ν^- ; l'équation différentielle correspondante est

$$\frac{dy}{ds} = \nu(y) \quad ; y(0) = (0,a) \qquad (8)$$

La solution $y(s)$ admet Γ_a comme trajectoire et s correspond à la longueur d'arc puisque $|\nu(y)| = 1$. Un de ces choix

dessinera l'a avec s positif, et l'autre avec s négatif.

Soit v_0 le vecteur unité de $K(0,a)$ pour lequel la première coordonnée (λ) est positive. Si $v^+(0,a) = v^0$, prenons $v(y) = v^+(y)$ et $v^-(y)$ sinon. D'où on a choisi une équation différentielle particulière dans $\mathcal{O}$.

Le système (8) peut s'écrire (par définition de $K(\lambda,x)$)

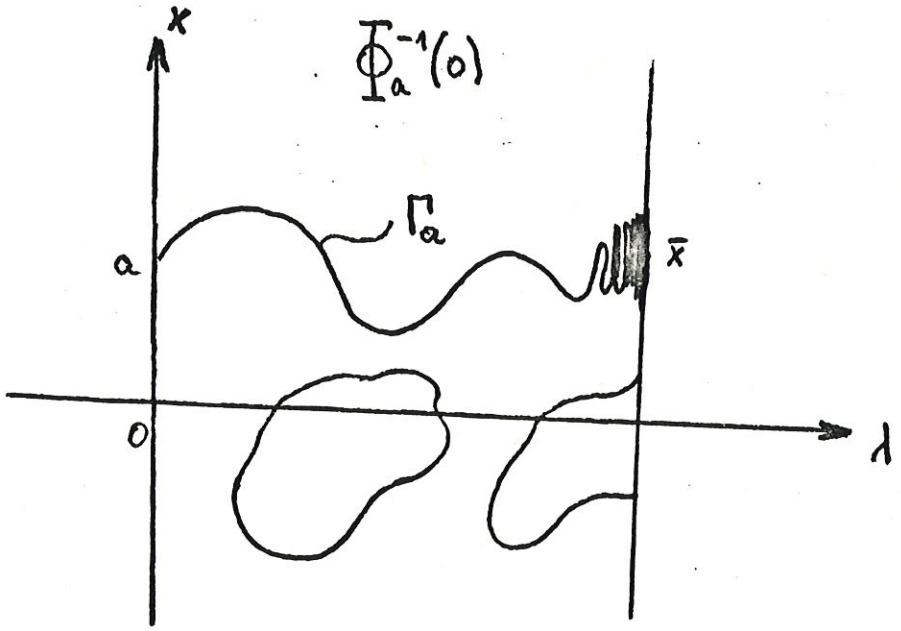
$$\begin{cases} D\Phi_a(\lambda,x)\begin{pmatrix} \frac{d\lambda}{ds} \\ \frac{dx}{ds} \end{pmatrix} = 0 & x(0) = a \\ \left\| \left(\frac{d\lambda}{ds}, \frac{dx}{ds} \right) \right\| = 1 & \lambda(0) = 0 \end{cases} \qquad (9)$$

ce qui représente le départ du problème de programmation qu'on traitera en détail au chapitre suivant.

Dans le cas où $I - \lambda Df(x)$ est singulière au point fixe, (9) peut avoir une infinité de solutions près de ce point. Ceci engendre par conséquent un ensemble dense de points fixes. Soit en effet $\bar{x}$ un tel point avec $\operatorname{rg}(I - Df(\bar{x})) = n-2$. Le système (9) devient $[a - f(x), I - \lambda Df(x)] \begin{bmatrix} \frac{d\lambda}{ds} \\ \frac{dx}{ds} \end{bmatrix} = 0$; $\left\| \left(\frac{d\lambda}{ds}, \frac{dx}{ds} \right) \right\| = 1$

avec les conditions initiales $\lambda(0)=0$; $x(0)=a$.

En $\lambda=1$, il y a une infinité de solution $\bar{x}$ qui engendre un ensemble dense de points fixes. La courbe Γ_a se mettra à osciller près de $\lambda=1$ et ne sera par conséquent plus de longueur finie (fig. 12).



CHAPITRE III :

L'algorithme du point fixe

Le chapitre II vient de donner les éléments théoriques pour construire un algorithme cherchant des points fixes d'applications régulières de $\bar{B}^n$ vers $\bar{B}^n$. Dans ce paragraphe nous allons décrire en détail cet algorithme et fournir quelques résultats numériques. On se base surtout sur l'article [8].

III.1. Position du problème

Soit une fonction $f : \bar{B}^n \longrightarrow \bar{B}^n$ de classe C^2. Les constatations faites au chapitre II restent aussi vraies pour le cas des ensembles compacts convexes généraux. On propose de chercher un point fixe de f.
Utilisons en plus $g_a : \bar{B}^n \longrightarrow \bar{B}^n$ une application C^2 dont on connaît une racine (en général, soit $g_a(x) = x - a$; la racine est donc $a \in \bar{B}^n$).
Soit l'application homotopique $\phi(\cdot,\cdot,\cdot) : \bar{B}^n \times [0,1] \times \bar{B}^n \longrightarrow \bar{B}^n$ telle que $\phi(a,\lambda,x) = \lambda(x - f(x)) + (1-\lambda) g_a(x)$. En plus, prenons $\phi_a(\lambda,x)$ comme étant égal par notation à $\phi(a,\lambda,x)$.

Comme vu au chapitre précédent, on cherchera les solutions (λ, x) de $\phi_a^{-1}(0)$ tout en parcourant λ de 0 à 1. En partant en $\lambda = 0$ avec $(0, a)$ on arrive en $(1, \bar{x})$ où $f(\bar{x}) = \bar{x}$ est solution du problème.

III.2. Convergence de l'algorithme

La justification théorique a été faite au chapitre II. Par le <u>théorème paramétrisé de Lord</u> on a le

Théorème III.2.1.

Soit $\phi(a, \lambda, x)$ une application C^2 de $\bar{B}^n \times [0,1] \times \bar{B}^n$ vers $\mathbb{R}^n$ telle que ϕ soit transversale à zéro.

Alors, pour presque tout $a \in \mathrm{Int}\,\bar{B}^n$ $\phi_a(\lambda, x) \equiv \phi(a, \lambda, x)$ est transversale à zéro c-à-d pour presque tout $a \in \mathrm{Int}\,\bar{B}^n$ la matrice jacobienne de $\phi_a(\lambda, x)$ est de rang plein en $\phi_a^{-1}(0)$.

Théorème III.2.2

Soit $\phi : \bar{B}^n \times [0,1] \times \bar{B}^n \longrightarrow \mathbb{R}^n$ telle que $\phi(a, \lambda, x) = \lambda(x - f(x)) + (1-\lambda)(x -$ où $x - f(x)$ admet un <u>jacobien non-singulier</u> en tout point fixe de f.

Alors, pour presque tout $a \in \mathrm{Int}\,\bar{B}^n$ l'ensemble $\{(\lambda, x) \mid 0 \le \lambda \le 1, x \in \bar{B}^n, \phi_a(\lambda, x) = 0\}$ de zéros de ϕ_a est constitué par

1) un nombre fini de courbes fermées (à longueur finie) dans $]0,1[\times \bar{B}^n$.

2) un nombre fini d'arcs (à longueur finie) dans $]0,1[\times \overline{B}^n$ à points d'arrivée en $\{1\} \times \overline{B}^n$.

3) une courbe à longueur finie partant de $(0,a)$ et arrivant en $(1,\bar{x})$ où $\bar{x} \in \overline{B}^n$ est un point fixe de f.

Les courbes 1), 2) et 3) sont disjointes et continuement différentiables.

Le théorème III.2.2. résulte du théorème II.3.2. et du fait que $|I - Df(x)|$ Il suffit de raisonner au moyen du théorème des fonctions implicites et d'utiliser le fait que $\phi_a^{-1}(0)$ est constitué par un ensemble de courbes régulières (Corollaire II.3.1.).

Sans demander trop d'hypothèses de différentiabilité, la non-singularité de la matrice jacobienne de $x - f(x)$ aux points fixes de f est nécessaire pour garantir une longueur finie des courbes nulles de $\phi_a(\lambda, x)$. En effet, il faut éviter une indétermination du système différentiel $D\phi_a \cdot \frac{dy}{ds} = 0$. Sinon, on aurait une oscillation infiniment grande de $\phi_a^{-1}(0)$ près de $\lambda = 1$ et que l'ensemble des solutions devienne dense sur une partie de $\overline{B}^n$.

L'algorithme pour la recherche de points fixes de f est extrêmement simple: en partant avec $\lambda = 0$, $a \in \operatorname{Int} \overline{B}^n$ (souvent $a \equiv$ origine) tout en suivant la courbe $\phi_a^{-1}(0)$ (ie sa composante Γ_a) on arrive (théorème III.2.2.3)) en $(1,\bar{x})$ où $\bar{x}$ est le point fixe de f. Cette courbe Γ_a est de longueur finie.

III.3. L'algorithme numérique

Construire un programme suivant la courbe nulle de $\phi_a^{-1}(0)$ n'est pas si facile qu'on pourrait le croire au début. La courbe ne doit pas être suivie de trop près (car ceci exige plus d'informations que sont nécessaires en réalité). Si la courbe est suivie de trop loin, on peut tomber sur une courbe voisine et retourner en arrière vers $\lambda = 0$ (cfr. plus loin).

III.3.1. Le programme et ses difficultés

Il faut suivre la courbe nulle de

$$\phi_a(\lambda, x) = \lambda\,(x - f(x)) + (1-\lambda)(x-a) \qquad (1)$$

en partant de $(0, a)$. En général, $a = 0_{\mathbb{R}^n}$.

La forme (1) est utilisée pour des méthodes continues pour lesquelles λ est une variable indépendante. Les méthodes échouent si la courbe retourne parce que x n'est pas défini de façon unique en fonction de λ (voir fig. 1).

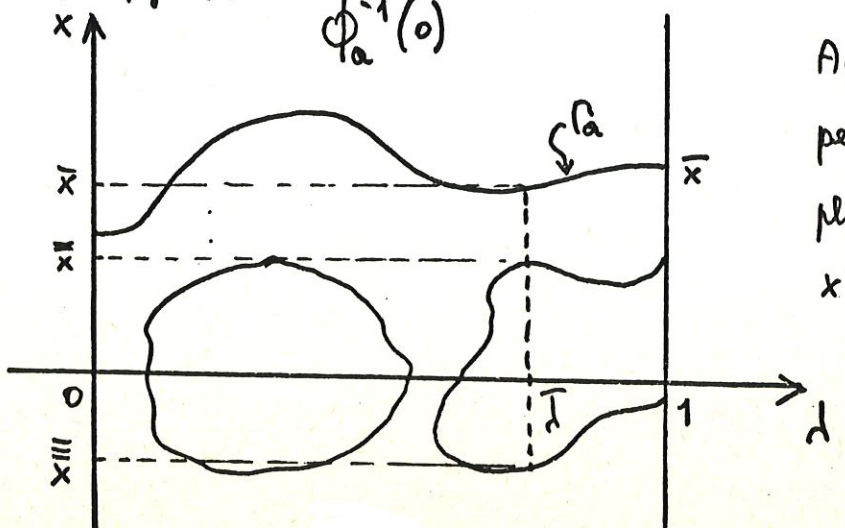


fig. 1

Au certain $\bar{\lambda}$ donné peuvent correspondre plusieurs valeurs de x : $\bar{x}$, $\bar{\bar{x}}$ et $\bar{\bar{\bar{x}}}$ par exemple

le problème sera évité dans notre algorithme en faisant de λ une variable dépendante. Un choix raisonnable pour la variable indépendante est la longueur d'arc (notée s). On obtient donc par (1) que

$$\lambda(s)\,(x(s)-f(x(s)))+(1-\lambda(s))\,(x(s)-a)=0 \qquad (2)$$

Cette courbe nulle de $\phi_a(\lambda,x)$ partant de $(0,a)$ est solution du problème à valeur initiale (voir II.5)

$$\begin{cases} \dfrac{d}{ds}\,\phi_a(\lambda(s),x(s))=0 & \lambda(0)=0 \\ \left\| \left(\dfrac{d\lambda}{ds},\dfrac{dx}{ds}\right)\right\|_2 = 1 & x(0)=a \end{cases} \qquad (3)$$

Si la solution $(\lambda(s),x(s))$ de (3) atteint $\lambda(s)=1$, alors le $x(s)$ correspondant est un point fixe de f.

Le calcul de $\left(\dfrac{d\lambda(s)}{ds},\dfrac{dx(s)}{ds}\right)$ est la partie la plus coûteuse de (3). Il sera donc nécessaire d'utiliser une intégration d'équations différentielles (ED) minimisant le plus possible le nombre d'évaluations. Par exemple, les subroutines STEP (résoudre l'ED) et INTRP (interpoler le point fixe) sont très satisfaisants dans ce contexte. On va les décrire dans ce chapitre.

Soit $y(s)\equiv(\lambda(s),x(s))$. Pour utiliser les subroutines d'intégration numérique d'ED standards, (3) doit être mis sous la forme $Y'(s)=G(s,Y)$. Ceci sera fait par les subroutines FODE et DCPOSE qui vont être spécifiées dans ce chapitre.

La première équation de (3) s'écrit

$$\frac{d}{ds}\phi_a(\lambda(s), x(s)) = 0$$

$$\Leftrightarrow \quad D\phi_a(\lambda(s), x(s))\ \frac{dy}{ds} = 0$$

$$\Leftrightarrow \quad \left[D_\lambda \phi_a(\lambda, x) \quad D_x \phi_a(\lambda, x) \right] \begin{bmatrix} \frac{d\lambda}{ds} \\ \frac{dx}{ds} \end{bmatrix} = 0$$

$$\Leftrightarrow \quad \left[a - f(x) \quad I - \lambda\, Df(x) \right] \begin{bmatrix} \frac{d\lambda}{ds} \\ \frac{dx}{ds} \end{bmatrix} = 0 \qquad (4)$$

où $Df(x)$ représente la matrice jacobienne de f.

Donc $y'(s)$ peut être trouvé en calculant le noyau d'une matrice $n \times n+1$ dont le rang est plein (car celui de $I - \lambda Df(x)$ l'est par hyp.)

Comme la solution de l'ED ne contrôle pas l'erreur locale, l'augmentation de la longueur d'arc de la courbe monte le danger que les points trouvés diffèrent de la courbe actuelle. C'est pour cela que si l'arc est trop long, le dernier point trouvé $(\bar\lambda, \bar x)$ ne sera utilisé pour calculer

$$\bar a = \frac{\bar x - \bar\lambda f(\bar x)}{1 - \bar\lambda} \qquad (5)$$

Comme $\phi_{\bar a}(\bar\lambda, \bar x) = 0 \Leftrightarrow (5)$, on a que le point $\bar a$ défini par (5) se trouve exactement sur la courbe nulle de $\phi_{\bar a}(\bar\lambda, \bar x)$.
On suit donc la courbe nulle de $\phi_{\bar a}(\bar\lambda, \bar x)$ en partant de $(\bar\lambda, \bar x)$. Si $\bar a \notin$ Int $\bar B^m$, la solution trouvée $(\bar\lambda, \bar x)$ est trop loin de la courbe cherchée il faudra suivre la courbe originale de plus près.

En passant de $\phi_a(\lambda, x)$ à $\phi_{\bar{a}}(\lambda, x)$ on risque de <u>perdre "le sens de direction"</u>. Les courbes ϕ_a et $\phi_{\bar{a}}$ peuvent être situées de façon à ce qu'elles évolueront dans la même direction le long de $\phi_a = 0$ et $\phi_{\bar{a}} = 0$ (mais en sens différent) en retournant vers $\bar{a}$ (voir fig. 2).

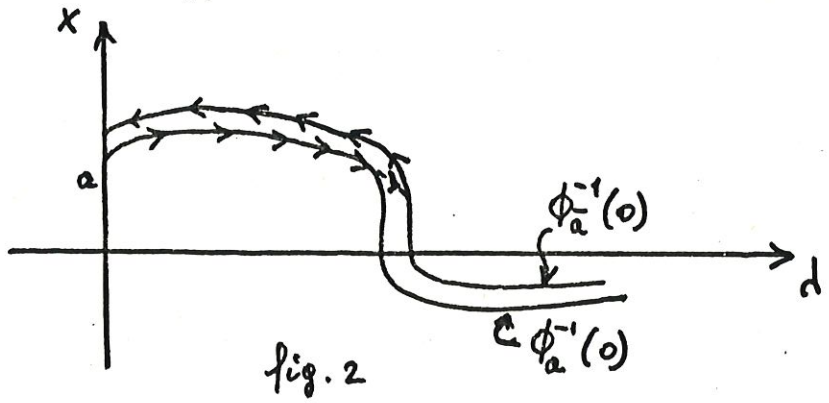


fig. 2

Ceci peut être évité en suivant la courbe de plus près, mais ceci augmente l'effort de programmation de façon considérable. Le danger le plus fort de "perdre" le sens de la courbe se présente lorsque la courbe tourne en arrière ou s'il y a un "virage" ou de façon équivalente : si $\frac{d\lambda}{ds}$ est petit. Des résultats numériques montrent qu'une bonne technique consiste à suivre la courbe de façon assez généreuse, sauf aux tournants où il faut rester aussi près que le permet la précision de la machine utilisée.

Comme le but donné est de trouver correctement un point fixe de f et non pas de résoudre (3) de façon exacte, il n'est pas indispensable de suivre la courbe nulle très correctement pour $\lambda \leq .99$. Quand λ atteint .99, on réinitialise le problème avec un nouveau vecteur

$\bar{a}$ (donné par (5)) de façon à ce que le dernier point calculé $(\lambda, \bar{x})$ se trouve exactement sur la courbe nulle de $\phi_{\bar{a}}(\lambda, x)$. On suit alors de façon précise la courbe $\phi_{\bar{a}}(\lambda, x)$ vers le point fixe cherché.

Un point fixe $x(\hat{s})$ de $f(x)$ correspond à $\lambda(\hat{s}) = 1$ et la valeur de $\hat{s}$ est inconnue avant (longueur d'arc de Γ_a). L'intégration numérique de (3) avance pas par pas et va donc un peu plus loin que $\hat{s}$ ce qui donnera un $\lambda(s) > 1$ (c'est la condition d'arrêt de l'intégration numérique). En utilisant une interpolation inverse l'équation $\lambda(s) = 1$ peut être résolue pour $\hat{s}$. Des points trouvés sur la courbe Γ_a près de $(\lambda(\hat{s}), x(\hat{s}))$ sont disponibles à partir de l'intégration de l'ED. Ils peuvent être interpolés pour trouver $(\lambda(s), x(s))$ pour tout s près de $\hat{s}$. Donc le calcul de $\hat{s}$ et $x(\hat{s})$ n'exige pas d'évaluations supplémentaires de fonctions. La subroutine ROOT cherche le $\hat{s}$ tel que $\lambda(\hat{s}) = 1$ et INTRP approxime de façon polynomiale la solution de x au point $\hat{s}$. Le résultat $x(\hat{s})$ sera le point fixe désiré.

L'intégration de (3) au moyen de STEP se base sur une méthode d'Adams. On pourrait suggérer de suivre la courbe Γ_a par la méthode d'Euler et de changer l'application (1) après chaque pas accordé à (5). Une autre possibilité est de prendre des pas donnés par Euler et de retourner vers la courbe exacte via Newton. Des résolutions numériques

faites par L.T. Watson dans [8] montrent cependant qu'aucune des deux méthodes n'est efficace comparée avec celle d'Adams.

L'algorithme du point fixe sera mis sous forme d'une subroutine appelée FIXPT :

#1 Choisir $a \in \operatorname{Int} \bar{B}^m$ (souvent $a = 0$).
Borne d'erreur locale ARCTOL > 0
Borne d'erreur finale $\equiv$ EPS > 0
En général, EPS $<<$ ARCTOL
Posons $s = 0$, $d(0) = 0$ et $x(0) = a$.

#2 Si la longueur d'arc est trop grande, aller en #3.
Si $d \geq .99$ et le problème n'a pas encore été posé avec la borne d'erreur finale EPS, poser la borne d'erreur égale à EPS et aller en #3. Sinon, aller en #4.

#3 Programmer un nouveau vecteur initial $\bar{a} = \dfrac{\bar{x} - \bar{I}\, f(\bar{x})}{1 - I}$ où $(I, \bar{x})$ est le point courant trouvé avant. Mettre des indications pour montrer le redépart de l'intégration de l'ED en se basant sur $\phi_{\bar{a}}(s, x)$.

#4 Faire un pas le long de la courbe en intégrant l'ED (on applique STEP).

#5 Si le pas échoue, aller en #10 (i)
Si on exige trop de précision, changer EPS ou ARCTOL ; aller en #10 (ii)
Si on exige trop peu de précision, " " " (ii)

#6 Mettre la borne d'erreur EPS si $\left|\frac{d\lambda}{ds}\right| < .01$ ou $\lambda \geq .99$
Sinon on garde la borne d'erreur ARCTOL pour suivre Γ_a.

#7 Si on a effectué trop d'évaluations de jacobiens, aller en #10.

#8 Si $\lambda < 1.$, aller en #2.

#9 Interpolation inverse pour chercher $\hat{s}$ tel que $\lambda(\hat{s}) = 1.$ Le $x(\hat{s})$ correspondant sera le point fixe cherché.

#10 Indicateurs appropriés et retour au programme principal.

IFLAG = 1 si pas de problème
= 2 si on vient de #5 (ii)
= 3 " #7
= 4 #5 (i)
= 5 " #5 (iii)

III.3.2. Les subroutines

Le programme principal se base essentiellement sur la subroutine FIXPT décrite précédemment. Dans le programme on ne fait appel qu'à FIXPT. Les autres subroutines sont utilisées dans FIXPT même. On va les décrire dans ce paragraphe. Le langage utilisé est le FORTRAN.

III.3.2.1. La subroutine FIXPT

Elle cherche les points fixes de la fonction donnée et représente la partie clef du programme principal. On l'appelle par

CALL FIXPT (N, Y, ARCTOL, EPS, ARCLEN, NFE, IFLAG)

où N = dimension du problème

Y = (d, x) vecteur de dimension N+1 contenant $(\bar{d}, \bar{x})$ en retour. A un retour normal, $\bar{d}=1$ et $\bar{x}$ est le point fixe de f

ARCTOL = erreur locale permise pour l'intégration numérique de l'ED. FIXPT automatiquement diminue cette erreur partout où la courbe est large, mais ARCTOL est utilisée partout où c'est possible. En un retour normal on a que ARCTOL = EPS.

EPS = précision totale désirée au point fixe calculé. En général EPS << ARCTOL. Un bon choix est ARCTOL = SQRT(EPS)

ARCLEN = longueur d'arc de la courbe suivie. En un retour sans succes ARCLEN peut être faux.

NFE = nombre d'évaluations de jacobiens. Le nombre d'évaluations de f est en général NFE+1 ou même plus grand. NFE est correct en un retour sans succes.

IFLAG = indice d'erreur

= 0 à l'appel de FIXPT

= 1 en cas d'un retour normal (où $\bar{d}=1$, $\bar{x}$ = point fix

= 2 indique que EPS (ou ARCTOL) est trop petit. ARCTOL et EPS ont été changés par FIXPT. La résolution du problème peut être continuée en appelant FIXPT sans changer aucun des paramètres.

= 3 indique qu'il y a un grand travail à effectuer ce qui veut dire que le problème est très difficile ou ARCTOL est trop petit. On peut continuer la résolution en appelant FIXPT avec IFLAG = 3.

= 4 indique qu'on est tombé sur un jacobien singulier : erreur fatale.

= 5 montre que la solution calculée est allée trop loin de Γ_a. On reprend le problème avec IFLAG=0 et un ARCTOL (et EPS) plus petit.

= 6 indique un jacobien singulier de $x - f(x)$ près du point fixe. La solution peut être douteuse.

N est un argument ; ARCTOL, EPS et IFLAG sont transitoires et Y, ARCLEN et NFE sont des résultats de FIXPT.

On utilise dans FIXPT les subroutines suivantes :

STEP résout le problème (3) par une intégration numérique d'ED au moyen d'une méthode d'Adams.

ROOT cherche un $\hat{s}$ tq $d(\hat{s}) = 1$ et INTRP calcule le $x(\hat{s})$ correspondant qui représente le point fixe de f.

FODE spécifie l'ED pour STEP. Elle calcule le noyau de

$D\phi_a(t,x)$ et utilise la subroutine DCPOSE pour ce travail.

F calcule f en un certain point et met le résultat dans un vecteur

FJAC évalue la k^e colonne du jacobien de f en un certain point et place le tout dans un vecteur.

III.3.2.2. La subroutine STEP

Elle résout une équation du premier ordre de la forme

$$\begin{cases} y'(x) = f(x, y(x)) \\ y(a) = y_0 \end{cases} \qquad (6)$$

STEP travaille avec un ordre variable et utilise une formule d'Adams en combinant une extrapolation avec un PECE (prédicteur d'ordre k et correcteur d'ordre $k+1$) pour résoudre (6).

Chaque appel à STEP fait avancer la résolution d'un pas (variable) de x_n vers $x_{n+1} = x_n + h$. L'erreur locale est contrôlée au moyen d'un critère d'erreur qui fait varier l'ordre et la grandeur du pas. En général, les changements d'ordre sont limités à un et la longueur du pas ne peut être divisée en deux ou doublée. L'ordre est limité à 12 et avec l'extrapolation sera limité à 13. La procédure part elle-même à partir des conditions initiales a et y_0 en commençant l'intégration à l'ordre 1.

L'appel à STEP se fait par

CALL STEP(X, Y, FODE, NEQN, H, EPS, WT, START, HOLD,
K, KOLD, CRASH, PHI, P, YP, PSI)

où X = variable indépendante mise à x_n (joué par la longueur d'arc s).

Y = Vecteur solution y_n en x_n (dans notre cas "y_n" = (λ, x))

FODE = Subroutine externe de la forme FODE(X, Y, YP). Elle transforme la matrice jacobienne de (3) et (4) de façon à avoir un système $y'(x) = G(x, y(x))$. Elle cherche alors le noyau de $D\phi_a(\lambda, x)$ dans (4) pour trouver $y'(x) = \left(\frac{d\lambda}{dx}, \frac{dx}{dx}\right)$ où "x" = s.

NEQN = Nombre d'équations de (3) (dans notre cas NEQN = N+1)

H = Longueur optimale du pas suivant.

WT(NEQN) = Poids pour le contrôle d'erreur.

START = variable logique qui vaut .TRUE. pour le premier pas et .FALSE. sinon.

HOLD = longueur du pas précédent

K = ordre optimal pour le pas suivant.

KOLD = ordre du dernier pas fait.

CRASH = variable logique qui vaut .TRUE. si le pas ne peut être fait et .FALSE. sinon.

PHI(NEQN, 16) = Matrice de changement des différences divisées. Les colonnes 15 et 16 servent au contrôle.

P(NEQN) = solution provisoire en x_n.

YP(NEQN) = dérivée de la solution corrigée en x_n.

PSI(12) = coefficients $\psi_i(n) = h_n + h_{n-1} + \dots + h_{n-i+1}$.

III.3.2.3. La subroutine ROOT

La subroutine ROOT(T,FT,B,C,RELERR,ABSERR,IFLAG) cherche une racine de l'équation non-linéaire F(T)=0.
On utilise un intervalle [B,C] tel que F(B).F(C)<0. En supposant que F soit continue dans cet intervalle, il y aura nécessairement une racine entre B et C. Des approximations successives vers la racine sont effectuées par une méthode de sécante en diminuant l'intervalle [B,C] contenant la racine. Le critère d'arrêt consiste à avoir $\left|\frac{B-C}{2}\right| \leq |B| * \text{RELERR} + \text{ABSERR}$ où F(B) et F(C) sont de signes opposés. B est une meilleure approximation au sens que |F(B)| ≤ |F(C)|. Les quantités ABSERR et RELERR sont les erreurs absolues et relatives. La valeur F(T) est demandée à chaque approximation T. Pour l'obtenir, le code retourne avec un IFLAG < 0. Le calcul FT = F(T) se fait souvent par interpolation (INTRP).
ROOT continue à diminuer l'intervalle jusqu'au moment de convergence ou jusqu'au cas de conditions contradictoires. En tout cas, on retourne avec un IFLAG > 0. IFLAG doit être positif au premier appel de ROOT pour effectuer l'initialisation. Si on retourne avec IFLAG négatif, on calcule (au moyen de INTRP) FT = F(T) et on fait de nouveau appel à ROOT. Si IFLAG > 0, on a terminé.

Les paramètres T, FT, B, C et IFLAG sont des variables du programme appelant (elles sont utilisées pour "l'input" et "l'output").

Les différentes valeurs possibles pour IFLAG sont :

IFLAG = 1 Si F(B)·F(C) < 0 et le critère d'arrêt est réalisé.

= 2 Si une valeur B est trouvée telle que F(B) est exactement zéro. L'intervalle (B, C) ne pourra satisfaire le critère d'arrêt : on a trop de précision.

= 3 Si |F(B)| dépasse les valeurs entrées de |F(B)| et |F(C)|. Dans ce cas B est près d'un pôle de f.

= 4 On ne trouve pas de racines dans l'intervalle. Il se peut qu'on soit tombé sur un minimum local.

= 5 On a effectué trop d'évaluations de fonctions.

Dans notre programme (FIXPT) ROOT doit chercher un $\hat{s}$ tel que $d(\hat{s}) =$

```
      ⋮
  170 CALL ROOT (SOUT, Y1SOUT, SA, SB, EPS, EPS, LCODE)
      IF (LCODE .GT. 0) GOTO 190
      CALL INTRP (S, Y, SOUT, WT, P, NP1, KOLD, PHI, PSI)
      Y1SOUT = WT(1) - 1.0
      GOTO 170
  190 IFLAG = 1   (Sortie normale)
      ⋮
```

Chaque INTRP calcule un vecteur WT = (d, X) de façon à ce que "FT" = Y1SOUT soit égal à WT(1) - 1.0 = d - 1.0. Ainsi (en déterminant T tel que FT = 0), on détermine la valeur $\hat{s}$ (se trouvant en "B" = SA) telle que $d(\hat{s}) = 1$.

III.3.2.4. La subroutine INTRP

La subroutine INTRP (X, Y, XOUT, YPOUT, NEQN, KOLD, PHI, PSI) calcule le vecteur solution YOUT et son vecteur dérivé YPOUT au point "outpout" XOUT.
On approxime la solution près de X par un polynôme. X se trouve près de XOUT c-à-d $|X - HOLD| < |XOUT| \leq |x|$. L'information pour définir ce polynôme est passée par STEP, donc INTRP ne peut être utilisé seul. Certains des paramètres sont repris de STEP : NEQN, KOLD, PHI, PSI, X, Y.

Les subroutines STEP, ROOT et INTRP sont expliquées dans [9].

III.3.2.5. La subroutine FODE

Cette subroutine résout la première équation du système

$$\begin{cases} [a - f(x), I - \lambda Df(x)] \begin{bmatrix} \frac{d\lambda}{ds} \\ \frac{dx}{ds} \end{bmatrix} = 0 \\ \left\| \left(\frac{d\lambda}{ds}, \frac{dx}{ds} \right) \right\|_2 = 1 \end{cases} \qquad (3)$$

Ceci revient à chercher le noyau de la matrice $D\phi_a(\lambda, x)$ qui est par hypothèse de rang plein n. FODE met le problème sous la forme

$$Y'(s) = G(s, Y) \quad \text{où} \quad \begin{cases} Y = (\lambda, x) \\ x = x(s) \\ \lambda = \lambda(s) \end{cases} \qquad (7)$$

Pour cela on considère la matrice

$$QR \equiv D\phi_a(\lambda, x) = [a - f(x),\ I - \lambda Df(x)]$$

qu'on va réduire sous forme triangulaire supérieure au moyen de la subroutine DCPOSE. Ceci permet de résoudre (7).

L'appel à FODE se fait dans STEP même (FODE est un des paramètres de STEP). Il sera de la forme

CALL FODE(S, Y, YP)

où S = longueur d'arc (paramètre indépendant)

Y = solution (λ, x) de (3)

YP = vecteur solution dérivé c-à-d $YP = Y' = \frac{dY}{ds}$

(solution de (7)).

La subroutine est tirée de [11] que vous pouvez consulter pour plus d'informations.

III.3.2.6. La subroutine DCPOSE

Elle transforme la matrice jacobienne de (3) à une forme triangulaire supérieure. Ceci se fait par pivotage en utilisant des des arguments de moindres carrés trouvés par P. Businger. Pour plus d'informations voir [10]

L'appel se fait par

CALL DCPOSE (NDIM, N, QR, ALPHA, PIVOT, IERR, SUM)

où NDIM = nombre maximal de lignes de QR

N = dimension du problème

QR = Matrice jacobienne de $\phi_a(\lambda, x)$. A l'appel elle vaut $[a - f(x), I - \lambda Df(x)]$ et sera transformée à forme triangulaire supérieure au retour.

ALPHA = Vecteur contenant les éléments diagonaux de la nouvelle matrice QR.

PIVOT = Vecteur contenant les indices colonnes des éléments pris comme pivots.

SUM = Vecteur contenant la somme des carrés des éléments des différentes colonnes.

IERR = 0 pas de problème

= 1 matrice singulière.

Dans le cas IERR = 1, il faut arrêter le problème, car on tombe sur une matrice singulière et la méthode ne marche pas.

III.3.2.7 Les subroutines F, FJAC

- La subroutine F(X, V, N) place les valeurs de $f(x)$ dans le vecteur V lorsque le problème est de dimension N.
- La subroutine FJAC(X, V, K, N) calcule le jacobien de $f(x)$ par rapport à la variable X(K) (on reprend la K$^{\text{ème}}$ colonne de la matrice jacobienne $Df(x)$) et place le résultat dans le vecteur V.

III.3.3. Recherche de racines

La subroutine FIXPT peut être modifiée pour chercher les racines d'une fonction f. Pour cela, il suffit de se baser sur l'application homotopique $\phi_a(\lambda, x) = \lambda f(x) + (1-\lambda)(x-a)$. Il faut donc modifier la matrice jacobienne $D\phi_a(\lambda, x)$ qui sera alors $[a - x + f(x), (1-\lambda) I + \lambda Df(x)]$

Le problème va être traité au chapitre suivant où on examine des problèmes d'optimisation. Les théorèmes de convergences indispensables s'y trouvent aussi.

L'algorithme ne se limitera donc plus nécessairement à la boule $\overline{B}^n$. A part le fait qu'on n'a pas toujours la convergence de l'algorithme il y a deux changements importants par rapport aux programmes décrits précédemment :

a) La matrice jacobienne $D\phi_a(\lambda, x)$ sera égale à $[a - x + f(x), (1-\lambda) I + \lambda Df(x)]$. Les changements doivent être faits dans FODE.

b) Pour se ramener exactement sur la courbe Γ_a, il suffit de prendre $\bar{a} \equiv \dfrac{\bar{\lambda} f(\bar{x}) + (1-\bar{\lambda}) \bar{x}}{1 - \bar{\lambda}}$ ce qui équivaut à avoir $\phi_{\bar{a}}(\lambda, x) = 0$. Le changement sera fait dans FIXPT lui-même.

III.4. Résultats numériques

On cherche les points fixes d'applications C^2. L'ordinateur utilisé est le DEC 2020.

Rappelons que NFE = nombre d'évaluations de jacobiens.
ARCLEN = longueur de la courbe Γ_a.
N = dimension du problème.
EPS = précision finale
ARCTOL = précision locale
X = point fixe cherché

On travaille avec EPS = 1.0E-07 et ARCTOL = 1.0E-02 et $a = 0$

Les exemples II.4.1 à II.4.4. ont été suggérés par L.T. Watson dans [8]

III.4.1.

Soit $f : \bar{B}^n \longrightarrow \bar{B}^n$ telle que $f_k(x) = \frac{1}{2 \cdot N}\left(\sum_{i=1}^{N} x(i)^3 + k\right)$ $k = 1 \dots N$

On cherchera donc les vecteurs x tels que

$$f_k(x) = \frac{1}{2 \cdot N}\left(\sum_{i=1}^{N} x(i)^3 + k\right) = x(k) \quad k = 1 \dots N$$

N = 1 NFE = 43
ARCLEN = 0.11845E+01
X = 0.61803E+00

N = 2 NFE = 44
ARCLEN = 0.11807E+01
X = 0.29766E+00 0.54766E+00

N = 5
NFE = 36
ARCLEN = 0.12806E+01
X =

0.12736E+00	0.22736E+00	0.32736E+00	0.42736E+00	0.52736E+00

N = 10
NFE = 48
ARCLEN = 0.14472E+01
X =

0.72344E-01	0.12234E+00	0.17234E+00	0.22234E+00	0.27234E+00
0.32234E+00	0.37234E+00	0.42234E+00	0.47234E+00	0.52234E+00

N = 20
NFE = 44
ARCLEN = 0.17373E+01
X =

0.45093E-01	0.70093E-01	0.95093E-01	0.12009E+00	0.14509E+00
0.17009E+00	0.19509E+00	0.22009E+00	0.24509E+00	0.27009E+00
0.29509E+00	0.32009E+00	0.34509E+00	0.37009E+00	0.39509E+00
0.42009E+00	0.44509E+00	0.47009E+00	0.49509E+00	0.52009E+00

N = 30
NFE = 53
ARCLEN = 0.19861E+01
X =

0.36045E-01	0.52711E-01	0.69378E-01	0.86045E-01	0.10271E+00
0.11938E+00	0.13604E+00	0.15271E+00	0.16938E+00	0.18604E+00
0.20271E+00	0.21938E+00	0.23604E+00	0.25271E+00	0.26938E+00
0.28604E+00	0.30271E+00	0.31938E+00	0.33604E+00	0.35271E+00
0.36938E+00	0.38604E+00	0.40271E+00	0.41938E+00	0.43604E+00
0.45271E+00	0.46938E+00	0.48604E+00	0.50271E+00	0.51938E+00

N = 50
NFE = 69
ARCLEN = 0.24080E+01
X =

0.28818E-01	0.38818E-01	0.48818E-01	0.58818E-01	0.68818E-01
0.78818E-01	0.88818E-01	0.98818E-01	0.10882E+00	0.11882E+00
0.12882E+00	0.13882E+00	0.14882E+00	0.15882E+00	0.16882E+00
0.17882E+00	0.18882E+00	0.19882E+00	0.20882E+00	0.21882E+00
0.22882E+00	0.23882E+00	0.24882E+00	0.25882E+00	0.26882E+00
0.27882E+00	0.28882E+00	0.29882E+00	0.30882E+00	0.31882E+00
0.32882E+00	0.33882E+00	0.34882E+00	0.35882E+00	0.36882E+00
0.37882E+00	0.38882E+00	0.39882E+00	0.40882E+00	0.41882E+00
0.42882E+00	0.43882E+00	0.44882E+00	0.45882E+00	0.46882E+00
0.47882E+00	0.48882E+00	0.49882E+00	0.50882E+00	0.51882E+00

III.4.2. Soit $f : \overline{B}^n \to \overline{B}^n$ telle que $f_K(x) = .01 \left(\sum_{i=K-1}^{K+1} x(i) + 1 \right)^3 \quad K = 2 \ldots N-1$

$f_1(x) = .01 \left(x(1) + x(2) + 1 \right)^3$

$f_N(x) = .01 \left(x(N-1) + x(N) + 1 \right)^3$

N = 1
NFE = 36
ARCLEN = 0.10001E+01
X = 0.10313E-01

N = 2
NFE = 36
ARCLEN = 0.10001E+01
X = 0.10653E-01 0.10653E-01

N = 5
NFE = 36
ARCLEN = 0.10003E+01
X = 0.10655E-01 0.11014E-01 0.11025E-01 0.11014E-01 0.10655E-01

N = 10
NFE = 36
ARCLEN = 0.10006E+01
X = x(1) = x(10) = 0.10655E-01
x(2) = x(9) = 0.11014E-01
x(I) = 0.11025E-01 I = 3 ... 8

N = 30
NFE = 36
ARCLEN = 0.10018E+01
X = x(1) = x(30) = 0.10655E-01
x(2) = x(29) = 0.11014E-01
x(I) = 0.11025E-01 I = 3 ... 28

N = 50
NFE = 36
ARCLEN = 0.10030E+01
X = x(1) = x(50) = 0.10655E-01 x(2) = x(49) = 0.11014E-01
x(I) = 0.11025E-01 I = 3 ... 48

III.4.3. Soit $f : \mathbb{R}^n_{\mathbb{C}} \longrightarrow \mathbb{R}^n_{\mathbb{C}}$ c-à-d $f : C \to C$ telle que pour $k = 1 \ldots$

$$f_k(x) = \exp\left(\cos\left(k \sum_{i=1}^{N} x(i)\right)\right) \quad \text{où } C = [-e, e]^n$$

On a une application d'un compact convexe de $\mathbb{R}^n$ dans lui-même ce qui permet donc bien d'appliquer l'algorithme du point fixe.

C'est un exemple extrêmement difficile ; il a été trouvé par Watson en 1979. La difficulté provient du fait que le paramètre d (qui n'en est pas en réalité) change de façon globalement croissante c-à-d qu'il monte finalement à la valeur 1, mais qu'il y a beaucoup de passages où il diminue, puis augmente et diminue de nouveau ... Cette attitude de d augmente la complexité de l'a. Elle se manifeste par un nombre très élevé d'évaluations de jacobiens et une longueur d'arc très grande ce qui augmente l'effort de calcul de façon très considérable. L'a aura une allure graphique complète comme le montre la figure 3.

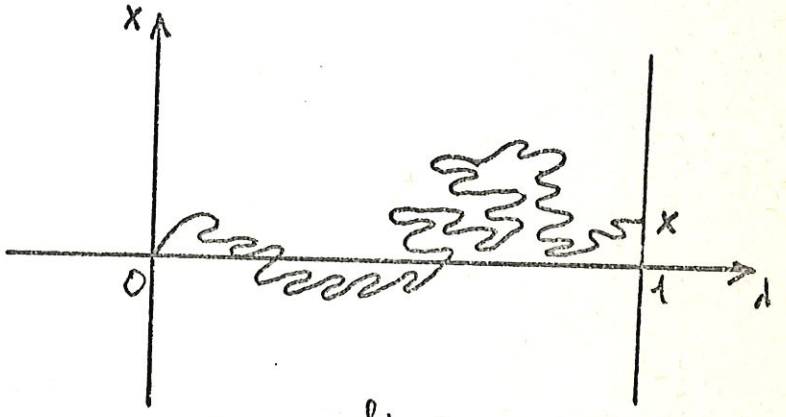


fig. 3

N=1
NFE = 79
RRCLEN = 0.18867E+01
X = 0.13030E+01

N=2
NFE = 89
RRCLEN = 0.16199E+01
X = 0.11004E+01 0.37467E+00

N=3
NFE = 309 RRCLEN = 0.51125E+01
X = 0.37473E+00 0.25267E+01 0.43255E+00

N=5
NFE= 812
RRCLEN= 0.14828E+02
X= 0.15876E+01 0.56399E+00 0.37096E+00 0.70894E+00 0.19614E+01

N=8
NFE= 1269
RRCLEN= 0.48678E+02
X= 0.39859E+00 0.19980E+01 0.70225E+00 0.95897E+00 0.15381E+00
0.47230E+00 0.25843E+01 0.36917E+00

N=10
NFE=7990
RRCLEN= 0.87504E+02
X= 0.14919E+00 0.50677E+00 0.38904E+00 0.92732E+00 0.24198E+01
0.21870E+00 0.77292E+00 0.37209E+00 0.58659E+00 0.17538E+01

III.4.4. : Soit $f: \mathbb{R}^n \longrightarrow \mathbb{R}^n$ telle que

$$f_1(x) = x(1) - \left[\prod_{i=1}^{n} x(i) - 1\right]$$

$$f_K(x) = x(K) - \left[\sum_{i=1}^{N} x(i) - x(K) - (N+1)\right] \qquad K = 2 \ldots N$$

C'est un exemple très difficile trouvé par Watson en 1978. Ceci vient du fait que la matrice jacobienne est très mal conditionnée au point fixe (elle est presque singulière). Cela risque de provoquer un manque de précision des résultats. La courbe Γa peut se mettre à osciller près du point fixe ce qui entraîne une augmentation de la longueur d'arc (voir fig. 4)

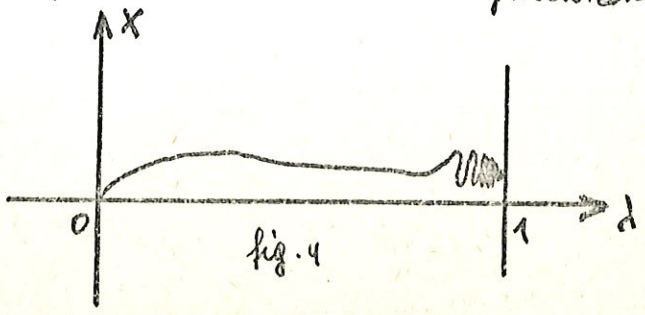


fig. 4

Pour n'importe quelle dimension on trouve $X(I) = 0.10000E+01$, $I = 1 \ldots N$.

N=1 NFE = 49 ARCLEN = 0.14412 E+01

N=2 NFE = 51 ARCLEN = 0.18279 E+01

N=5 NFE = 73 ARCLEN = 0.27118 E+01

N=10 NFE = 196 ARCLEN = 0.37199 E+01

N=20 NFE = 211 ARCLEN = 0.51259 E+01

N=30 NFE = 131 ARCLEN = 0.61888 E+01

III.4.5. Soit $f : \bar{B}^n \longrightarrow \bar{B}^m$ telle que

$$f_K(x) = \left(\frac{1}{100 \cdot N}\right) \cdot \left(EXP\left(2 \cdot \sum_{i=1}^{N} x(i)\right) + 2\,x(K) - 3 \cdot K\right)$$

$K = 1 \ldots N$

On trouve les résultats suivants :

N=1
NFE = 42
ARCLEN = 0.10002 E+01
X = -0.20824E-01

N=2
NFE = 42
ARCLEN = 0.10004E+01
X = -0.10452E-01 -0.25604E-01

N=5
NFE = 42
ARCLEN = 0.10009E+01
X =

-0.43192E-02 -0.10343E-01 -0.16367E-01 -0.22392E-01 -0.28416E-0

N=10
NFE = 42
ARCLEN = 0.10016E+01
X =

-0.22755E-02 -0.52815E-02 -0.82876E-02 -0.11294E-01 -0.14300E-0
-0.17306E-01 -0.20312E-01 -0.23318E-01 -0.26324E-01 -0.29330E-0

N= 30

NFE= 48

ARCLEN= 0.10074E+01

X=

-0.86809E-03	-0.18688E-02	-0.28694E-02	-0.38701E-02	-0.48708E-02
-0.58714E-02	-0.68721E-02	-0.78728E-02	-0.88734E-02	-0.98741E-02
-0.10875E-01	-0.11875E-01	-0.12876E-01	-0.13877E-01	-0.14877E-01
-0.15878E-01	-0.16879E-01	-0.17879E-01	-0.18880E-01	-0.19881E-01
-0.20881E-01	-0.21882E-01	-0.22883E-01	-0.23883E-01	-0.24884E-01
-0.25885E-01	-0.26885E-01	-0.27886E-01	-0.28887E-01	-0.29887E-01

N= 50

NFE= 50

ARCLEN= 0.10077E+01

X=

-0.55675E-03	-0.11570E-02	-0.17572E-02	-0.23575E-02	-0.29577E-02
-0.35580E-02	-0.41582E-02	-0.47584E-02	-0.53587E-02	-0.59589E-02
-0.65592E-02	-0.71594E-02	-0.77596E-02	-0.83599E-02	-0.89601E-02
-0.95604E-02	-0.10161E-01	-0.10761E-01	-0.11361E-01	-0.11961E-01
-0.12562E-01	-0.13162E-01	-0.13762E-01	-0.14362E-01	-0.14963E-01
-0.15563E-01	-0.16163E-01	-0.16763E-01	-0.17363E-01	-0.17964E-01
-0.18564E-01	-0.19164E-01	-0.19764E-01	-0.20365E-01	-0.20965E-01
-0.21565E-01	-0.22165E-01	-0.22766E-01	-0.23366E-01	-0.23966E-01
-0.24566E-01	-0.25167E-01	-0.25767E-01	-0.26367E-01	-0.26967E-01
-0.27568E-01	-0.28168E-01	-0.28768E-01	-0.29368E-01	-0.29969E-01

Remarque : Dans l'exemple III.4.5. on a cherché un x tel que

$$f_K(x) = x(K) \qquad K = 1 \ldots N$$

$$\Leftrightarrow EXP\left(2 \cdot \sum_{i=1}^{N} x(i)\right) - (100 \cdot N - 2) \cdot x(K) - 3 \cdot K = 0$$

$$\Leftrightarrow f_K^*(x) = 0 \qquad \text{où } f_K^* = f_K + x(K)$$

On constate qu'il est en pratique plus facile de résoudre $f_K(x) = x(K)$ au moyen de $\lambda(x - f(x)) + (1-\lambda)(x-a)$ que de calculer $(f_K(x) + x(K)) = f_K^*(x) = 0$ avec l'homotopie $\lambda f^*(x) + (1-\lambda)(x-a)$. Ceci résulte du fait que l'algorithme est préis surtout pour la recherche de points fixes.

Le seul problème est d'avoir des fonctions à racines dans $\overline{B}^n$ ou dans un compact bien défini pour savoir appliquer l'algorithme du point fixe.

CHAPITRE IV :

Applications en optimisation

Nous allons résoudre dans ce chapitre des problèmes d'optimisation au moyen de l'algorithme décrit dans le paragraphe précédent. Les applications étudiées ont été proposées par L.T. Watson dans [13]. On examinera des problèmes de minimisation sans et avec contraintes que l'on transformera en un problème de recherche de racines.

Soit $f : \mathbb{R}^n \to \mathbb{R}^n$ de classe C^2.
Définissons l'application homotopique $\phi_a(\lambda, x) = \lambda f(x) + (1-\lambda)(x-a)$

Lemme IV.0.

Pour presque tout $a \in \mathbb{R}^n$, il y a une courbe nulle Γ_a de $\phi_a(\lambda, x)$, partant de $(0, a)$, suivant laquelle $D\phi_a(\lambda, x)$ est de rang plein. En plus, Γ_a arrive en $\bar{x}$ (racine de f) ou va à l'infini. Si Γ_a arrive en $\bar{x}$ et la matrice jacobienne $Df(\bar{x})$ est non-singulière alors Γ_a est de longueur finie.

Le lemme résulte du corollaire II.3.1 de même façon que le théorème II.3.2. en utilisant la remarque II.3.4.2. En se basant sur la remarque II.3.4.1. on voit que $D\phi_a(\lambda,\bar{x}) = [a + f(\bar{x}) - \bar{x},\ Df(\bar{x})]$ sera de rang plein si $Df(\bar{x})$ est non-singulière ce qui entraîne une longueur d'arc finie de Γ_a. La remarque II.3.4.3. reste évidemment valable.

IV.1. Minimisation sans contraintes

On propose dans la suite de minimiser une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$ qui est de classe C^3 (c-à-d $\nabla f(\cdot)$ de classe C^2).

Théorème IV.1.1.

Soit $f : \mathbb{R}^n \to \mathbb{R}$ une application convexe de classe C^3 où minimum en $\bar{x}$; $\|\bar{x}\| \leq M$.

Alors : pour presque tout $a \in \mathbb{R}^n$, $\|a\| < M$, il y a une courbe nulle Γ_a de l'application homotopique $\phi_a(\lambda,x) = \lambda \nabla f(x) + (1-\lambda)(x-a)$ suivant laquelle $D\phi_a(\lambda,x)$ est de rang plein. Cette courbe relie $(0,a)$ à $(1,\bar{x})$ où $f(\bar{x}) = \min_{x \in \mathbb{R}^n} f(x)$. En plus, si le Hessien $H(\bar{x}) = D(\nabla f(\bar{x}))$ est non singulier, Γ_a sera de longueur finie.

Preuve : Comme $D\phi_a(\lambda, x) = \left[Df(x) - x + a,\ (1-\lambda) I + \lambda D^2 f(x) \right]$, il suit que $D\phi_a(1, \bar{x}) = \left[Df(\bar{x}) - \bar{x} + a,\ D^2 f(\bar{x}) \right]$. (*)

- L'existence de Γ_a et le fait que $D\phi_a(\lambda, x)$ est de rang plein le long de Γ_a résulte du lemme IV.0. : il suffit de remarquer que f est C^3 entraîne que Df est de classe C^2 et Df va de $\mathbb{R}^n$ vers $\mathbb{R}^n$.
- Par (*), on voit que par l'application du lemme IV.0., la non-singularité de $D^2 f(\bar{x})$ entraîne une longueur finie de Γ_a.
- Pour montrer que Γ_a atteint un zéro de Df (ie un minimum de f) il suffit de vérifier que f est borné (ce qui évite le cas infini).

Soit (λ, x), $0 \le \lambda < 1$, un point quelconque de Γ_a; $\|x\| = 3M$. Comme $\|a\| < M$ et $\|\tilde{x}\| \le M$, on a que $(x - \tilde{x})(x - a) > 0$; d'où $(1-\lambda)(x - \tilde{x})(x - a) \ge 0$ (i). Par la convexité de f, $f(x) \ge f(\tilde{x}) + Df(\tilde{x})(x - \tilde{x})$ ou encore que $f(\tilde{x}) - f(x) \ge Df(x)(\tilde{x} - x)$. Comme $\tilde{x}$ est un minimum de f, on déduit que $f(\tilde{x}) - f(x) \le 0$. En plus $(x - \tilde{x}) Df(x) = (x - \tilde{x})(Df(x) - Df(\tilde{x})) \ge 0$. Ce qui entraîne que $(x - \tilde{x})\, \lambda\, Df(x) \ge 0$ pour $0 \le \lambda < 1$ (ii). L'inégalité (i) implique $(x - \tilde{x}) \left[\lambda Df(x) + (1-\lambda)(x - a) \right] > 0$. D'où, $\phi_a(\lambda, x) \neq 0$ pour $0 \le \lambda < 1$ et $\|x\| = 3M$. Γ_a sera contenu dans $[0,1] \times \{x \mid \|x\| \le 3M\}$, d'où borné.

■

Le théorème permet donc d'appliquer l'algorithme décrit au chapitre III adapté à la recherche des racines de Df : ceci résout notre problème.

IV.2. Minimisation avec contraintes de positivité : l'aspect théorique

Nous traitons dans la suite le problème d'optimisation suivant :

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} f(x) \qquad (1) \\ &\text{sc} \quad x \geq 0 \end{aligned}$$

où $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ est une application convexe de classe C^3.

La première partie de ce paragraphe montre que le problème (1) est équivalent au problème de complémentarité

$$\begin{cases} x \geq 0 \\ F(x) \geq 0 \\ x \cdot F(x) = 0 \end{cases} \qquad (2) \qquad \text{où } F(x) \equiv \nabla f(x)$$

c-à-d $F_i(x) = \dfrac{\partial f(x)}{\partial x_i}$ $i = 1 \dots n$

La seconde partie va établir l'équivalence entre (2) et le système d'équations non-linéaires

$$(|F_i(x) - x(i)|)^3 - (F_i(x))^3 - (x(i))^3 = 0 \qquad (3)$$

$$i = 1 \dots n.$$

La résolution du problème (1) revient donc à chercher la racine de $G(x)$ où $G_i(x) = (|F_i(x) - x(i)|)^3 - (F_i(x))^3 - (x(i))^3$, $i = 1 \dots n$

Le passage de (1) à (2) est basé sur [14] (théorèmes 7.2.1 et 7.3.7.) tandis que l'équivalence de (2) et (3) utilise les réflexions de [15].

IV.2.1. Les conditions de Kuhn - Tucker

On essayera dans la suite de traiter le problème (1) de façon plus générale et d'établir l'équivalence avec le problème de complémentarité (2).

Soit donc le problème de minimisation (PM)

$$f(\bar{x}) = \min_{x \in X} f(x) \quad \text{où } X = \{ x \in X^0 \mid g(x) \leq 0 \}. \qquad (1')$$

$g : \mathbb{R}^n \to \mathbb{R}^m$; X^0 ouvert de X

qu'on peut même examiner de façon locale ce qui me donnera (PML)

$$f(\bar{x}) = \min_{x \in X} f(x) \quad \text{tel que } x \in B_\delta(\bar{x}) \cap X \Rightarrow f(x) \geqslant f(\bar{x}) \qquad (1'')$$

$\delta > 0$.

En prenant $g(x) \equiv -x$, on obtient l'équivalence de (1) et (1').

IV.2.1.1. Définitions

On appelle <u>conditions de Kuhn - Tucker</u> (KTP) le système suivant.

Chercher un $\bar{x} \in X^0$, $\bar{u} \in \mathbb{R}^m$ tels que

$$\left\{ \begin{array}{l} \nabla_x \Psi(\bar{x}, \bar{u}) = 0 \\ \nabla_u \Psi(\bar{x}, \bar{u}) \leq 0 \\ \bar{u}\, \nabla_u \Psi(\bar{x}, \bar{u}) = 0 \\ \bar{u} \geqslant 0 \\ \text{où } \Psi(x, u) = f(x) + u\, g(x) \end{array} \right\} \text{ ou de façon équivalente } \left\{ \begin{array}{ll} \nabla f(\bar{x}) + \bar{u} \nabla g(\bar{x}) = 0 & \text{(i)} \\ g(\bar{x}) \leq 0 & \text{(ii)} \\ \bar{u}\, g(\bar{x}) = 0 & \text{(iii)} \\ \bar{u} \geqslant 0 & \text{(iv)} \end{array} \right.$$

Le système KTP correspond bien au problème de complémentarité (2).

En effet, le fait que $g(x) \equiv -x$ et (ii) entraînent $\bar{x} \geq 0$. Or, $\nabla g(\bar{x}) = -Id$
d'où par (i) on a que $\nabla f(\bar{x}) = \bar{u}$ et donc par (iv) $\nabla f(\bar{x}) \geq 0$. Finalement
la condition (iii) implique que $\bar{u} \cdot g(\bar{x}) = 0$ ssi $\nabla f(\bar{x}) \cdot g(\bar{x}) = 0$ ssi $\nabla f(\bar{x}) \cdot \bar{x} = 0$

Pour montrer l'équivalence entre (1) et (2), il suffit de la prouver entre
(1') et KTP dans le cas $g(x) = -x$.

IV.2.1.2. Conditions suffisantes d'optimalité : (1') $\Leftarrow$ KTP

Théorème IV.2.1.2

Soit $\bar{x} \in X$, X^0 ouvert de X contenant $\bar{x}$.
Soient f et g différentiables et convexes en $\bar{x}$.
Si $(\bar{x}, \bar{u})$ est solution de KTP, alors $\bar{x}$ sera solution de PNL.

Preuve: Soit $(\bar{x}, \bar{u})$ solution de KTP. Par convexité et différentiabilité
de f on a que pour tout $x \in X$ $f(x) - f(\bar{x}) \geq \nabla f(\bar{x})(x - \bar{x})$. Comme
$\nabla f(\bar{x}) = -\bar{u} \nabla g(\bar{x})$, il suit que $f(x) - f(\bar{x}) \geq -\bar{u} \nabla g(\bar{x})(x - \bar{x}) \geq \bar{u}[g(\bar{x}) - g(x)]$
par convexité et différentiabilité de g en $\bar{x}$. Le fait que $\bar{u}\, g(\bar{x}) = 0$ et que
$\bar{u} \geq 0$ et $g(x) \leq 0$ entraînent que $f(x) - f(\bar{x}) \geq -\bar{u}\, g(x) \geq 0$. D'où $\forall x \in$
$f(x) \geq f(\bar{x})$. Puisque $g(\bar{x}) \leq 0$, $\bar{x} \in X$ on a que $f(\bar{x}) = \min_{x \in X} f(x)$ et $\bar{x} \in X$. ∎

La convexité de $g(x)$ implique qu'on ne peut pas traiter des contraintes
égalités non-linéaires $h(x) = 0$ en les transformant en $h(x) \leq 0$ et $-h(x) \leq 0$.
Le théorème précédent illustre le passage de KTP à (1') et donc de (2) à (1).

IV.2.1.3. Conditions nécessaires d'optimalité : (1') $\Rightarrow$ KTP

1. La qualification de contraintes de Kuhn-Tucker (QC)

Soit X^0 ouvert de $\mathbb{R}^n$; $g : X^0 \to \mathbb{R}^m$; $I = \{ i \mid g_i(\bar{x}) = 0 \}$.

Alors, g satisfait la qualification de contraintes de Kuhn-Tucker en $\bar{x}$

ssi - g est différentiable en $\bar{x}$

- s'il existe $y \in \mathbb{R}^n$ tel que $(\nabla g_i(\bar{x}) \cdot y)_{i \in I} \leq 0$, alors il existe $e : [0,1] \to \mathbb{R}^n$ tel que

  a. $e(0) = \bar{x}$

  b. $e(z) \in X$ $0 \leq z \leq 1$

  c. e est différentiable en $z = 0$ et $\frac{de(0)}{dz} = \lambda y$ pour $\lambda > 0$

Le problème (1) satisfait bien (QC). En effet, $\nabla g(\bar{x}) \cdot y = -Iy = -y \leq 0$ c-à-d $y \geq 0$ entraîne qu'il existe $e : [0,1] \to \mathbb{R}^n$ tel que $e(z) \equiv zy + (-z^2)\bar{x}$ vérifiant les conditions a. $e(0) = \bar{x}$

b. $e(z) \in \mathbb{R}^n$ pour tout $z \in [0,1]$.

c. $\frac{de}{dz}\Big|_{z=0} = y - 2z\bar{x}\Big|_{z=0} = y$ $(\lambda = 1 > 0)$.

2. Le théorème de Kuhn-Tucker

Admettons d'abord le théorème suivant. Il sera utile dans la suite ; pour une preuve détaillée voir [14, p.28].

Théorème de l'alternative (Motzkin)

Soient A, C et D des matrices données

Alors : soit (I) : $Ax > 0$, $Cx \geq 0$, $Dx = 0$ a une solution x.

soit (II) : $\left\langle \begin{matrix} A'y_1 + C'y_3 + D'y_4 = 0 \\ y_1 \geq 0 \ ; y_3 \geq 0 \end{matrix} \right\rangle$ a une solution y_1, y_3, y_4

mais jamais (I) et (II) en même temps.

Théorème IV.2.1.3. (Kuhn-Tucker)

Soit X° ouvert de $\mathbb{R}^n$; f, g différentiables en $\bar{x}$ et définies sur X°.
Soit $\bar{x}$ solution de (PNL) ou (PNLL) ; g satisfait (CQC) en $\bar{x}$.
Alors, il existe un $\bar{u} \in \mathbb{R}^m$ tel que $(\bar{x}, \bar{u})$ soit solution de KTP.

Preuve : Soit $\bar{x}$ solution du (PNL) avec $\delta = \bar{\delta}$; $I = \{ i \mid g_i(\bar{x}) = 0 \}$ et $J = \{ i \mid g_i(\bar{x}) < 0 \}$. On doit traiter les cas $I = \phi$ et $I \neq \phi$.

a) Si $I = \phi$. Prenons un $y \in \mathbb{R}^n$ quelconque tel que $y'y = 1$. Alors on a que $g_i(\bar{x} + \delta y) = g_i(\bar{x}) + \delta \left[\nabla g_i(\bar{x})\, y + \alpha_i(\bar{x}, \delta y) \right]$, $i = 1 \dots m$. Comme $g_i(\bar{x}) < 0$ et $\lim_{\delta \to 0} \alpha_i(\bar{x}, \delta y) = 0$, on aura pour δ assez petit i.e. $0 < \delta < \hat{\delta} < \bar{\delta}$ que $g_i(\bar{x} + \delta y) < 0$ et $\bar{x} + \delta y \in X^\circ$ puisque X° est ouvert. Le fait que $\bar{x}$ résout (PNL) implique $0 \leq f(\bar{x} + \delta y) - f(\bar{x}) = \delta \left[\nabla f(\bar{x})\, y + \alpha(\bar{x}, \delta y) \right]$ pour $0 < \delta < \hat{\delta}$. D'où, $\nabla f(\bar{x})\, y + \alpha(\bar{x}, \delta y) \geq 0$. Puisque $\lim_{\delta \to 0} \alpha(\bar{x}, \delta y) = 0$ il suit que $\nabla f(\bar{x})\, y \geq 0$ pour δ assez petit. $y \in \mathbb{R}^n$ étant choisi de

façon quelconque tel que $y'y = 1$; on peut le prendre égal à $\pm e^i \in \mathbb{R}^n$ où e^i est le $i^{\text{ème}}$ vecteur unité. Il suit que $\nabla f(\bar{x}) = 0$. Donc $\bar{x}$ et $\bar{u} = 0$ satisfont bien KTP.

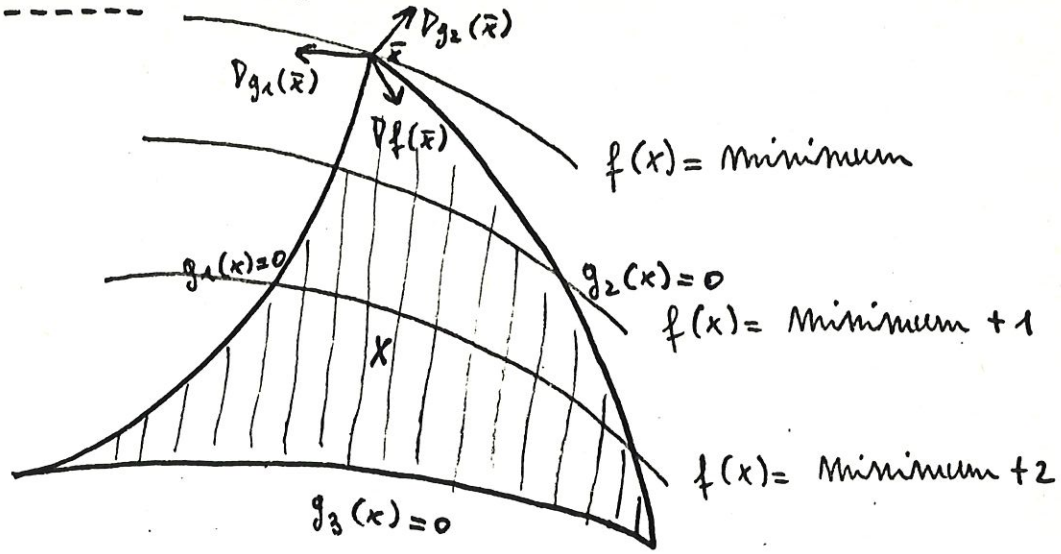
b) $I \neq \phi$. Prenons g satisfaisant (QC) en $\bar{x}$. Soit un $y \in \mathbb{R}^n$ tel que $(\nabla g_i(x) \cdot y)_{i \in I} \leq 0$. Il existe par (QC) un $e(\cdot) : [0,1] \to \mathbb{R}^n$ tel que $e(0) = \bar{x}$, $e(\tau) \in X$ pour $0 \leq \tau \leq 1$ et $\frac{de(0)}{d\tau} = \delta y$ pour $\delta > 0$. D'où pour $0 \leq \tau \leq 1$, $e_i(\tau) = e_i(0) + \tau \left[\frac{de_i(0)}{d\tau} + f_i(0,\tau) \right]$, $i = 1 \dots n$; où $\lim_{\tau \to 0} f_i(0,\tau) = 0$.

En prenant τ assez petit, $0 < \tau < \hat{\tau} < 1$, on aura $e(\tau) \in B_\delta(\bar{x})$. Comme $e(\tau) \in X$ pour $0 \leq \tau \leq 1$ et $\bar{x}$ résout (PNL) on déduit que $f[e(\tau)] \geq f[e(0)]$, $0 < \tau < \hat{\tau}$. Par la différentiabilité de f en $\bar{x}$ et e en 0, on a pour $0 < \tau < \hat{\tau}$ que $0 \leq f[e(\tau)] - f[e(0)] = \nabla f(e(0)) \frac{de(0)}{d\tau} \cdot \tau + \beta(0,\tau) \cdot \tau$ où $\lim_{\tau \to 0} \beta(0,\tau) = 0$. Ceci entraîne que $\nabla f[e(0)] \frac{de(0)}{d\tau} + \beta(0,\tau) \geq 0$ pour $0 < \tau < \hat{\tau}$ et pour τ assez petit $\nabla f(e(0)) \frac{de(0)}{d\tau} \geq 0$. Or $e(0) = \bar{x}$ et $\frac{de(0)}{d\tau} = \delta y$, $\delta > 0$ implique que $\nabla f(\bar{x}) \cdot y \geq 0$. On vient de voir que $(\nabla g_i(\bar{x}) y)_{i \in I} \leq 0$ entraîne $\nabla f(\bar{x}) \cdot y \geq 0$ c-à-d que $\left\langle \begin{matrix} \nabla f(x) y < 0 \\ (\nabla g_i(x) y)_{i \in I} \leq 0 \end{matrix} \right\rangle$ n'a pas de solution $y \in \mathbb{R}^n$. Par le thm de Motzkin il existe $\bar{\pi}_0$ et $\bar{\pi}_I$ tels que $\bar{\pi}_0 \nabla f(\bar{x}) + \bar{\pi}_I \nabla g_i(x) = 0$, $i \in I$; $\bar{\pi}_0 \geq 0$; $\bar{\pi}_I \geq 0$. Définissons $\bar{u}_I \equiv \frac{\bar{\pi}_I}{\bar{\pi}_0}$, $\bar{u}_J = 0$ et $\bar{u} = (\bar{u}_I, \bar{u}_J)$. Alors, $\nabla f(\bar{x}) + \bar{u} \nabla g(\bar{x}) = 0$; en plus $\bar{u} g(\bar{x}) = \sum_I \bar{u}_I \underbrace{g(\bar{x}_i)}_{=0} + \sum_J \underbrace{\bar{u}_J}_{=0} g(\bar{x}_j) = 0$ et $\bar{u} \geq 0$. Comme $\bar{x} \in X$, on déduit par définition de X que $g(\bar{x}) \leq 0$. Donc, $(\bar{x}, \bar{u})$ résout KTP.

■

Le théorème établit le passage de (1") à (KTP) c-à-d de (1) à (2).

Remarque : Interprétation des conditions de Kuhn-Tucker (KTP).



$\bar{x}$ est le seul point tel que $\nabla f(\bar{x}) + \sum_{i=1}^{3} \bar{u_i} \nabla g_i(\bar{x}) = 0$

Si on a que : $\bar{u} \geq 0$, $\bar{u} = (1,1,0)$

$\bar{u}\, g(\bar{x}) = 0 \Rightarrow$ $\bar{u_1} > 0$ $\quad \bar{g_1}(\bar{x}) = 0$
$\bar{u_2} > 0$ $\quad \bar{g_2}(\bar{x}) = 0$
$\bar{u_3} = 0$ $\quad \bar{g_3}(\bar{x}) < 0$

IV. 2.2. Systèmes d'équations non-linéaires

Le paragraphe va étudier l'équivalence entre le problème de complémentarité (2) et le système d'équations non-linéaires (3). En tenant compte du paragraphe IV.2.1., on aura donc ramené le problème (1) au système (3).

Théorème IV.2.2.1.

Soit θ une fonction strictement croissante de $\mathbb{R}$ dans $\mathbb{R}$ et soit $\theta(0)=0$.

Alors, x résout le problème de complémentarité (2)

ssi $\theta\left(\left|\frac{\partial L}{\partial x(i)}(x)-x(i)\right|\right)-\theta\left(\frac{\partial L}{\partial x(i)}(x)\right)-\theta(x(i))=0 \quad i=1\dots m \quad (2')$

Preuve: La condition est nécessaire. Pour tout $i=1\dots m$, $x_i=0$ ou $F_i(x)=0$ $(x_i \overset{not}{=} x(i))$. En effet, $x\cdot F(x)=\sum_{i=1}^{m} x_i F_i(x)=0$ et $x_i\geq 0$ et $F_i(x)\geq 0$. Si $x_i=0$, alors $\theta(|F_i(x)-x_i|)-\theta(F_i(x))-\theta(x_i)=\theta(|F(x_i)|)-\theta(F_i(x))$ $=\theta(F_i(x))-\theta(F_i(x))=0$. Si $F_i(x)=0$, alors $\theta(|F_i(x)-x_i|)-\theta(F_i(x))-\theta(x_i)$ $=\theta(|x_i|)-\theta(x_i)=\theta(x_i)-\theta(x_i)=0$

La condition est suffisante. (a) Voyons que $F(x)\geq 0$. Sinon il existe i tel que $F_i(x)<0$, d'où $0\leq\theta(|x_i-F_i(x)|)\overset{(2')}{=}\theta(F_i(x))+\theta(x_i)<\theta(x_i)$ puisque θ est strictement croissante et $\theta(0)=0$. Donc $x_i>0$ et en plus $x_i>|x_i-F_i(x)|=x_i-F_i(x)$ ce qui contradit $F_i(x)<0$. (b) Pour voir que $x\geq 0$, il suffit de changer les rôles de x et $F(x)$ et d'appliquer (a). (c) On a que $x\cdot F(x)=0$. Sinon il existe un i tel que $x_i>0$ et $F_i(x)>0$. Si $F_i(x)\geq x_i$, alors $\theta(|F_i(x)-x_i|)=\theta(F_i(x)-x_i)<\theta(F_i(x))$ $<\theta(F_i(x))+\theta(x_i)$ ce qui contradit le fait que par (2') $\theta(|F_i(x)-x_i|)$ $=\theta(F_i(x))+\theta(x_i)$. Si $x_i\geq F_i(x)$, alors $\theta(|x_i-F_i(x)|)=\theta(x_i-F_i(x))$ $<\theta(x_i)<\theta(x_i)+\theta(F_i(x))$ ce qui contradit également (2'). ∎

Pour l'application de l'algorithme utilisant $\phi_\alpha(\lambda,x)=\lambda G(x)+(1-\lambda)(x-a)$ où G est défini par (2'), il faut la non-singularité du jacobien de G.

Le corollaire suivant donne une condition suffisante pour que le jacobien de (2') soit non-singulier.

Corollaire IV.2.2.2.

Soit x la solution du problème de complémentarité (2) tel que $x + F(x) > 0$

Soit $\nabla F(x)$ le jacobien de F en x ayant des mineurs principaux non-singuliers. Soit θ une fonction strictement croissante, différentiable de $\mathbb{R}$ dans $\mathbb{R}$ telle que $\theta'(0) + \theta'(\xi) > 0$, $\forall \xi > 0$.

Alors, x résout (2') et la matrice jacobienne de (2') est non-singulière.

Preuve: (2') est équivalent à $\theta(|F_i(x) - x_i|) - \theta(F_i(x)) - \theta(x_i) = 0$, $i = 1 \dots n$,

ce qui donne par définition $G_i(x) = 0$, $i = 1 \dots n$.

Soit $\operatorname{sgn} \xi \equiv \begin{cases} 1 \text{ si } \xi \geq 0 \\ -1 \text{ si } \xi < 0 \end{cases}$, δ_{ij} le symbôle de Kroenecker.

Alors, $$\frac{\partial G_i(x)}{\partial x_j} = \theta'(|F_i(x) - x_i|) \operatorname{sgn}(F_i(x) - x_i) \left(\frac{\partial F_i(x)}{\partial x_j} - \delta_{ij} \right) - \theta'(F_i(x)) \frac{\partial F_i(x)}{\partial x_j} - \theta'(x_i) \delta_{ij}.$$

Supposons pour le moment que $F_i(x) = 0$, pour tout $i = 1 \dots \bar{m} \leq n$ et $F_i(x) > 0$ pour $i = \bar{m}+1 \dots n$. D'où, comme $x_i + F_i(x) > 0$ pour $i = 1 \dots n$ il suit que $x_i > 0$ pour $i = 1 \dots \bar{m}$ et $x_i = 0$ pour $i = \bar{m}+1 \dots \bar{n}$ et

$$\nabla G(x) = \begin{bmatrix} -\theta'(x_1) - \theta'(0) & & & & \\ & \ddots & & & \\ & & -\theta'(x_{\bar{m}}) - \theta'(0) & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \nabla F(x) + \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & & -\theta'(F_{\bar{m}+1}(x)) - \theta'(0) & \\ & & & & \ddots \\ & & & & & -\theta'(F_n(x)) - \theta'(0) \end{bmatrix}$$

La non-singularité de $\nabla G(x)$ suit du fait que $\theta'(0)+\theta'(\xi)>0$ pour $\xi>0$ et par la non-singularité des mineurs principaux $\frac{\partial F_i(x)}{\partial x_i}$ $i,j=1\dots\bar{m}$.

L'argumentation est similaire pour le cas $F_i(x)=0$ pour $i\in I\subset\{1\dots m\}$ et $I\neq\{1\dots\bar{m}\}$. ∎

La réalisation la plus facile pour (2') serait de prendre $\theta(x)=x$. Ceci me donne $|F_i(x)-x_i|-F_i(x)-x_i=0$, $i=1\dots m$. Le système admet bien un jacobien non-singulier sous les hypothèses du corollaire, mais il est seulement localement différentiable près d'une solution non-dégénérée. Ceci provient du fait que la fonction "la valeur absolue de" est non différentiable en 0. Le système sera donc non-différentiable si $F_i(x)-x_i$. Notons que la solution satisfaisant la condition de non-dégénérescence $F_i(x)-x_i$ est égale à $F_i(x)>0$ ou $x_i>0$.

Pour avoir une différentiabilité globale de (2'), on exige $\theta'(0)=0$. La fonction la plus simple vérifiant cela (et étant strict. croissante) est $\theta(x)=x^3$. (2') prendra donc la forme

$$(|F_i(x)-x_i|)^3-(F_i(x))^3-(x_i)^3=0 \qquad i=1\dots m \quad (3)$$

Le système résout (par le thm IV.2.2.1) aussi le problème de complémentarité (2) et donc le problème de départ (1).

IV.3. Résolution du problème de minimisation avec contraintes de positivité

IV.3.1. Position du problème

Pour résoudre le problème de minimisation (1) le problème de complémentarité (2), on est finalement arrivé à chercher la solution de

$$\left(\left|\frac{\partial f}{\partial x_i}(x) - x_i\right|\right)^3 - \left(\frac{\partial f}{\partial x_i}(x)\right)^3 - (x_i)^3 = 0 \qquad (3)$$

$$c - \grave{a} - d \qquad G(x) = 0$$

On traite souvent des problèmes linéaires de complémentarité $c-\grave{a}-d$ où $\nabla f(x) = F(x) = Mx + q$ où $q \in \mathbb{R}^n$ et $M \in L(\mathbb{R}^n)$. Le travail se fait de façon assez facile si une solution locale est connue. Or, le cas de convergence globale est plus difficile. C'est pour cela qu'on utilisera l'algorithme décrit au chapitre précédent.

IV.3.2. Le cas de dimension 1

Pour motiver le cas général, considérons le problème linéaire de complémentarité suivant : $x \geq 0$; $Mx + q \geq 0$; $x(Mx+q) = 0$; $x, M, q \in \mathbb{R}$.
On x résout le problème si il résout

$$G(x) = 0$$

où $G(x) = |Mx + q - x|^3 - (Mx+q)^3 - x^3$

Considérons l'application homotopique $\phi_a(\lambda, x) = \lambda G(x) + (1-\lambda)(x-a)$.

L'idée sera donc de tracer la courbe nulle de $\phi_a(\lambda, x)$ en partant de $(0, a)$ tout en espérant d'atteindre un zéro $\bar{x}$ de $G(x)$ (en $\lambda = 1$) après avoir parcouru une distance finie.

L'équation paramétrique $\lambda = \lambda(s)$; $x = x(s)$ de la courbe nulle, où s est la longueur d'arc, est donnée par (cfr. chapitre III)

$$\left[G(x) - x + a \;,\; (1-\lambda) I + \lambda DG(x) \right] \begin{bmatrix} \frac{d\lambda}{ds} \\ \frac{dx}{ds} \end{bmatrix} = 0 \qquad (4)$$

$$\lambda(0) = 0 \;;\; x(0) = a$$

Prenons par exemple $M = 1$ et $q = -1$, alors

$D\phi_a(\lambda, x) = \left[1 - (x-1)^3 - x^3 - x + a \;,\; 1 - \lambda + \lambda(-3(x-1)^2 - 3x^2) \right]$. On constate que $\frac{d\lambda}{ds} = 0$ si $D_x \phi_a(\lambda, x) = 1 - \lambda(-6x^2 + 6x - 3) = 0$

c-à-d si $x = \frac{1 \pm \sqrt{(2-5\lambda)/3\lambda}}{2}$ (*).

En regardant les signes des quantités présentées, il est facile à voir que cette courbe (*) contient les points tournants ($\frac{d\lambda}{ds} = 0$) des courbes nulles de ϕ_a. Elle constitue une barrière entre $\lambda = 0$ et $\lambda = 1$. Si une courbe nulle (excepté celle où $a = 1$) franchit la barrière, elle diverge vers l'infini en x en diminuant λ de façon à ce que la barrière joue le rôle d'asymptote (voir fig. 1).

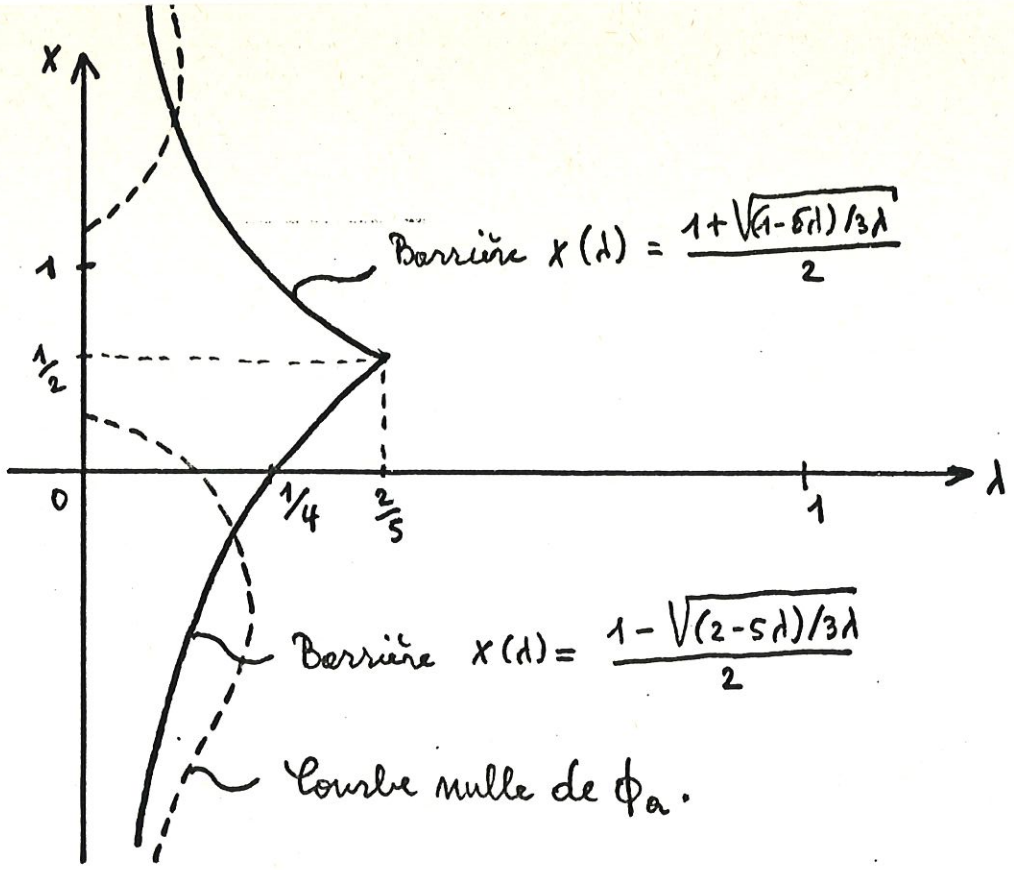


fig. 1

Observez que si l'homotopie utilise $-G(x)$ au lieu de $G(x)$, alors $D_x \phi_a(\lambda, x) > 0$ pour tout (λ, x), $0 \le \lambda \le 1$. Donc $\frac{d\lambda}{ds}$ ne s'annulle jamais ce qui équivaut à dire que la courbe nulle ne va pas retourner pour un a quelconque. Une analyse des signes des quantités variables montre que pour tout a la courbe nulle de ϕ_a atteint une solution $\bar{x} = 1$.

IV.3.3. Le cas général

Pour résoudre (3), soit $H(x) \equiv -G(x)$ et $\phi_a(\lambda, x) \equiv \lambda H(x) + (1-\lambda)(x-a$

L'utilité de ce changement est motivée par le paragraphe IV.3.2.

Nous appliquons maintenant notre algorithme décrit au chapitre III adapté à la recherche de racines dont la convergence sera établie

en utilisant surtout le raisonnement de L.T. Watson dans [16].

1) Convergence de l'algorithme

Une conséquence immédiate des théorèmes II.4.3. et II.3.2. est le

Lemme IV.3.3.1.

Soit $F : \mathbb{R}^n \to \mathbb{R}^n$ une application de classe C^2 telle que $x \cdot F(x) \geq 0$ pour tout x tel que $\|x\| = R$.

Alors, $F(x)$ admet un zéro dans la boule $\|x\| \leq R$ et pour presque tout a dans l'intérieur de la boule il y a une courbe nulle de l'application homotopique $\Psi_a(\lambda, x) = \lambda F(x) + (1-\lambda)(x-a)$ menant de $(0, a)$ vers un zéro de $F(x)$ le long de laquelle $rg\left(D\Psi_a(\lambda, x)\right) = n$.

Théorème IV.3.3.2.

Soit $F(x) = Mx + q$; $R = I$.

Alors, pour presque tout $a \in \mathbb{R}^n$, il y a une courbe nulle de $\phi_a(\lambda, x)$ menant de $(0, a)$ à $(0, \bar{x})$ où $\bar{x}$ résout le problème (linéaire) de complémentarité. $D\phi_a(\lambda, x)$ est de rang plein suivant cette courbe.

Preuve : On essayera de se mettre dans les conditions du lemme IV.3.3.

Soit $q_i = "F_i(x) - x_i"$; donc, $H_i(x) = -|q_i|^3 + (x_i + q_i)^3 + x_i^3$.

Ceci entraîne que $x \cdot H(x) = 2 \cdot \sum_{j=1}^{m} x_j^4 +$ termes d'ordre plus petit ≥ 0 si $\|x\| =$ suffisamment grand. La thèse suit alors du lemme IV.3.3.1. ∎

Le théorème de Sard (Thm II.1.2) peut s'exprimer sous la forme du

Lemme IV.3.3.3.

Soit $\phi : \mathbb{R}^n \times [0,1) \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ définie par $\phi(a,\lambda,x) = \lambda H(x) + (1-\lambda)(x-a)$. Alors, ϕ est transversale à zéro.

Le fait que 0 est valeur régulière de ϕ explique que $D\phi(a,\lambda,x)$ est de rang plein pour tout $(a,\lambda,x) \in \phi^{-1}(0)$; ceci n'est rien d'autre que la transversalité de ϕ en 0.

Le théorème paramétrisé de Sard (Thm II.1.3) est à la base du

Lemme IV.3.3.4.

Pour presque tout $a \in \mathbb{R}^n$, l'application $\phi_a : [0,1) \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ définie par $\phi_a(\lambda,x) = \lambda H(x) + (1-\lambda)(x-a)$ est transversale à 0 (c-à-d pour presque tout a, $D\phi_a(\lambda,x)$ est de rang plein en $\phi_a^{-1}(0)$).

Nous travaillons dans la suite avec cette application $\phi_a(\lambda,x)$. Remarquons que $H(x)$ est bien de classe C^2, $f(x)$ étant de classe C^3.

Lemme IV.3.3.5

Soit $DH(x)$ non-singulière en tout zéro de $H(x)$.

Alors, pour presque tout $a \in \mathbb{R}^n$, il y a une courbe nulle Γ_a de $\phi_a(\lambda, x)$ partant de $(0, a)$ suivant laquelle $D\phi_a(\lambda, x)$ est de rang plein. Γ_a est de longueur finie et atteint un zéro de $H(x)$ (en $\lambda = 1$) ou bien va à l'infini.

Preuve : L'existence de Γ_a et le rang plein de $D\phi_a(\lambda, x)$ sont tout simplement une conséquence du lemme IV.3.3.4. Supposons Γ_a bornée. Prolongeons ϕ_a et Γ_a sur $[0,1] \times \mathbb{R}^n$ suivant la définition donnée précédemment. Si $(\bar\lambda, \bar x)$ est un point de $[0,1] \times \mathbb{R}^n$, alors $D_x \phi_a(\bar\lambda, \bar x) = (1-\bar\lambda) I + \bar\lambda DH(\bar x)$ est de rang plein (car $D\phi_a$ l'est le long de la courbe et $DH(x)$ l'est en un point fixe). Par le théorème des fonctions implicites, on trouve un difféomorphisme entre Γ_a (en voisinage de $(\bar\lambda, \bar x)$) et un intervalle ouvert de $\mathbb{R}$ (voir la preuve du corollaire II.3.1). Il suit par la compacité de $[0,1]$ que Γ_a sera de longueur finie. En plus par le difféomorphisme entre Γ_a et un ouvert de $\mathbb{R}$, il est impossible que Γ_a se coupe avec elle même. Γ_a atteint $\lambda = 1$

Remarquons que Γ_a est une courbe C^1. Ceci suit du difféomorphisme avec des ouverts de $\mathbb{R}$.

Lemme IV.3.3.6.

Soient les hypothèses du lemme IV.3.3.5. Soit $a > 0$ tel que la conclusion de ce lemme soit satisfaite. Si $(\bar{\lambda}, \bar{x})$ se trouve sur la courbe nulle Γ_a de $\phi_a(\lambda, x)$ partant de $(0, a)$, alors $\bar{x} > 0$.

Preuve: La $i^{ème}$ composante de $\phi_a(\lambda, x)$ est donnée par

$$\lambda\left\{-|F_i(x)-x_i|^3+(F_i(x))^3+x_i^3\right\}+(1-\lambda)(x_i-a_i).$$

Supposons par l'absurde que $\bar{x}_i \leq 0$. Alors, $-|F_i(\bar{x})-\bar{x}_i|^3+(F_i(\bar{x}))^3+\bar{x}_i^3 \leq$ et $\bar{x}_i - a_i < 0$. Comme $0 \leq \bar{\lambda} < 1$, la i^e composante de $\phi_a(\lambda, x)$ est négative ce qui contredit le fait que $\phi_a(\bar{\lambda}, \bar{x}) = 0$. ∎

Théorème IV.3.3.7.

Soit la matrice jacobienne $DH(x)$ non-singulière en tout zéro de $H(x)$. Supposons qu'il existe un $r > 0$ tel que $x > 0$ et $x_a = \|x\|_\infty \geq r$ entraîne $F_a(x) >$ Alors, pour presque tout $a > 0$, il existe une courbe nulle Γ_a de $\phi_a(\lambda, x$

le long de laquelle $D\phi_a(\lambda, x)$ est de rang plein. Γ_a est de longueur finie et connecte $(0, \hat{a})$ à $(1, \bar{x})$ où $\bar{x} \equiv$ zéro de $H(x)$.

Preuve: L'existence de la courbe nulle Γ_a le long laquelle $D\phi_a(\lambda, x)$ est de rang plein suit du lemme IV.3.3.5. Si on parvient

à montrer que Γ_a est bornée, alors le théorème résulte du lemme IV.3.3.

Par le lemme IV.3.3.6, Γ_a se trouve dans $K = [0,1] \times \{x \in \mathbb{R}^n \mid x > 0\}$. Suppo
sons $r > \|a\|_\infty$. Soit $(\bar\lambda, \bar x) \in K$ un point tel que $\bar x_k = \|\bar x\|_\infty \geq r$.
Alors, $\bar x_k - a_k > 0$ et $-|F_k(\bar x) - \bar x_k|^3 + (F_k(\bar x))^3 + x_k^3 > 0$ (car $F_k(x) \geq 0$ et
$\bar x_k \geq r > \|a\|_\infty \geq a_k$). D'où, $\bar\lambda\{-|F_k(\bar x) - \bar x_k|^3 + (F_k(\bar x))^3 + \bar x_k^3\} + (1-\bar\lambda)(\bar x_k - a_k) >$
Ceci entraîne que $\phi_a(\lambda, x) \neq 0$ pour $0 \leq \lambda \leq 1$ et $\|x\|_\infty \geq r$. Il en suit
que Γ_a est contenu dans $[0,1] \times \{x \mid x \geq 0, \|x\|_\infty \leq r\}$, d'où borné. ∎

Le théorème IV.3.3.7 montre l'existence d'une solution du problème
de complémentarité (2) sous certaines conditions sur F, tout en se
basant sur l'algorithme décrit au chapitre III.

2) Le problème linéaire de complémentarité

Nous allons analyser le problème de complémentarité dans le cas $F = Mx + q$
où $q \in \mathbb{R}^n$ et M est une matrice $n \times n$.

Définitions IV.3.3.8.

1. M est strictement fortement dominante (SFD) ssi $|M_{ii}| > \sum_{j \neq i} |M_{ij}|$ $i = 1 \ldots$
2. M est définie positive (PD) ssi $x' M x > 0$ pour tout $x \neq 0$.
3. M est non-dégénérée (ND) ssi tous les mineurs principaux sont non-singul

4. M est une P-matrice (P) ssi tous les mineurs principaux sont positifs.
5. M est non-négative (NN) ssi chaque élément de M est non-négatif.
6. M est strictement copositive (SCP) ssi $x' M x > 0$ pour tout $x \geq 0$.
7. q est non-dégénéré par rapport à M ssi q n'est pas combinaison linéaire de $n-1$ ou moins de colonnes de $(I, -M)$.
8. M est strictement semi-monotone (SSM) ssi pour tout $x \geq 0, x \neq 0$, il existe un k tel que $x_k (Mx)_k > 0$.

Remarques : Il y a toute une série de relations entre ces définitions.

a. Illustrons la définition 7 : Le problème linéaire de complémentarité peut s'écrire
$$\begin{cases} w = Mx + q & \text{c-à-d } q = w - Mx \quad (i) \\ w \geq 0, \ x \geq 0, \ w^T x = 0 \end{cases}$$

Soit $L(q)$ l'espace déterminé par les contraintes $\begin{cases} w - Mx = q \\ w \in \mathbb{R}^n, \ x \in \mathbb{R}^n \end{cases}$ (ii)

Le vecteur $\begin{pmatrix} w \\ x \end{pmatrix} \in L(q)$ ssi il satisfait (ii).

q est non-dégénéré par rapport à M s'il ne se trouve pas dans un sous-espace engendré par $(n-1)$ ou moins colonnes de $(I, -M)$. Ceci entraîne que pour tout $(w, x) \in L(q)$, au plus n des $2n$ variables $\{w_j, x_j\}$ sont nulles.

Soit $\bar{w}, \bar{x}$ solution de (i) c-à-d $\bar{w} \geq 0, \bar{x} \geq 0$ et $\bar{w}^T \bar{x} = 0$. Il sera par conséquent évident que $\bar{w} + \bar{x} > 0$ c-à-d $\bar{x} + F(\bar{x}) > 0$.

b. Une matrice (SFD) à éléments diagonaux positifs est une P-matrice (voir [17]).

c. Une P-matrice est évidemment non-dégénérée (ND).

d. Une P-matrice est (SSM). Relation établie par M. Fiedler dans [18].

Corollaire IV.3.3.9.

La conclusion du théorème IV.3.3.7 reste valable dans le cas linéaire $F(x) = Mx+q$ où M est (SFD) à éléments diagonaux positifs et q est non-dégénéré par rapport à M.

Preuve: Soit $\bar{x}$ solution de (2). Comme q est non-dégénéré par rapport à M, il suit par la remarque a que $\bar{x} + F(\bar{x}) > 0$. En plus, $DF(\bar{x}) = M$ est (SFD) à éléments diagonaux positifs; d'où, par la remarque b, M est une P-matrice et sera par conséquent (ND). Par l'application du corollaire III.2.2.2., on conclut que $DH(x)$ est non-singulier en tout zéro de $H(x)$ (i). M étant (SFD) et $m_{ii} > 0$, on peut trouver un $r > 0$ tel que $M_{ii} - \sum_{j \neq i} |M_{ij}| + \frac{q_i}{r} > 0$. Soit $x_k = \|x\|_\infty \geq r$ et donc $|m_{kj}| \cdot \left|\frac{x_j}{x_k}\right| \leq |M_{kj}|$. On déduit que $F_k(x) = (Mx+q)_k = x_k \left(m_{kk} + \sum_{j \neq k} m_{kj} \frac{x_j}{x_k} + q_k / x_k\right) \geq x_k \left(m_{kk} - \sum_{j \neq k} |M_{kj}| + \frac{q_k}{x_k}\right) > 0$ (ii).

(i) et (ii) permet d'appliquer le théorème IV.3.3.7. D'où la thèse. ∎

Corollaire III.3.3.10.

La conclusion du thm III.3.3.7 reste valable dans le cas linéaire où $F(x) = Mx + q$ avec M une matrice (ND) et (NN) à éléments diagonaux positifs et q non dégénéré par rapport à M.

Preuve : Comme M est non-dégénérée, q non-dégénéré par rapport à M, on peut appliquer le corollaire III.2.2.2. D'où, $DH(x)$ non-singulière aux zéros de $H(x)$ (i). Choisissons un r tq $r > \frac{|q_i|}{M_{ii}} = \frac{|q_i|}{|M_{ii}|}$ pour $i = 1 \dots n$. Alors pour $x > 0$ et $x_k = \|x\|_\infty \geq r$ on a que $F_k(x)$ $= (Mx+q)_k = M_{kk} x_k + \sum_{j \neq k} M_{kj} x_j + q_k \geq M_{kk} r + q_k > 0$ (ii) (car $M_{kj} \geq 0$ pour $k, j = 1 \dots n$). (i) et (ii) permet d'appliquer le thm III.3.3 ce qui entraîne la thèse. ■

Pour des raisons techniques, il était plus raisonnable de définir ϕ_a sur $[0,1) \times \mathbb{R}^n$ sans traiter $\lambda = 1$. Dans la suite nous supposons pourtant que $\phi_a(\lambda, x)$ est défini sur $[0,1] \times \mathbb{R}^n$.

Lemme III.3.3.11.

Pour $a \geq 0$, tout zéro de $\phi_a(\lambda, x)$ satisfait $x \geq 0$.

Preuve : Supposons par l'absurde que $x_k < 0$ pour un $k \in \{1 \dots n\}$. Alors la k^e composante de $\phi_a(\lambda, x)$ satisfait
$\lambda \{ -|F_k(x) - x_k|^3 + (F_k(x))^3 + x_k^3 \} + (1-\lambda)(x_k - a_k) < 0$. Donc $\phi_a(\lambda, x) \neq 0$. ■

Lemme IV.3.3.12.

Supposons qu'il existe un $r>0$ tel que $x \geqslant 0$ et $\|x\|_\infty \geqslant r$ implique que $x_k F_k(x) > 0$ pour un certain indice k.

Alors, l'ensemble des zéros de $\phi_0(\lambda, x)$ est contenu dans $[0,1] \times \{x \geqslant 0 ; \|x\|_\infty < r\}$; d'où est borné.

Preuve : Soit $\phi_0(\bar\lambda, \bar x) = 0$. Par le lemme IV.3.3.11., il suit que $\bar x \geqslant 0$.

Si $\|\bar x\|_\infty \geqslant r > 0$, alors par hypothèse $x_k > 0$ et $F_k(x) > 0$ pour un certain k. Ceci implique que $(\phi_0(\bar\lambda, \bar x))_k = \bar\lambda \{ -|F_k(\bar x) - \bar x_k|^3 + (F_k(\bar x))^3 + x_k^3 \} + (1-\bar\lambda)\bar x_k > 0$ ce qui me donne une contradiction. Il en suit que $\|\bar x\|_\infty < r$ et d'où la thèse. ∎

Lemme IV.3.3.13.

Sous les hypothèses du lemme IV.3.3.12., il existe un $\delta > 0$ tel que $a \geqslant 0$ et $\|a\|_\infty < \delta$ entraîne $\phi_a(\lambda, x) \neq 0$ pour $0 \leqslant \lambda \leqslant 1$, $x \geqslant 0$ et $\|x\|_\infty$

Preuve : $\|\phi_0(\lambda, x)\|$ est une fonction continue sur le compact

$K = [0,1] \times \{x \mid x \geqslant 0 ; \|x\|_\infty = r\}$; elle admet donc une valeur minimale sur K. Par le lemme IV.3.3.12. $\min_K \|\phi_0(\lambda, x)\| = d > 0$

Soit $\Psi(a) = \min_K \|\phi_a(\lambda, x)\|$ une fonction continue en a. Comme $\Psi(0)$ on a que $\Psi(a) \neq 0$ au voisinage de 0 c-à-d si $\|a\|_\infty < \delta$. La conclusion du lemme vient du fait que $\Psi(a) \neq 0$ entraîne $\phi_a(\lambda, x) \neq 0$ sur K. ∎

Théorème IV.3.3.14

Soit $DH(x)$ non-singulier en tout zéro de $H(x)$. Supposons qu'il existe un $r > 0$ tel que $x \geq 0$ et $\|x\|_\infty > r$ entraine $x_k F_k(x) > 0$ pour un indice k.

Alors, il existe $\delta > 0$ tel que pour presque tout $a \geq 0$, $\|a\| < \delta$ il y a une courbe nulle Γ_a de $\phi_a(\lambda, x)$ le long de laquelle $D\phi_a(\lambda, x)$ est de rang plein, ayant une longueur finie et liant $(0, a)$ à $(1, \bar{x})$ où $\bar{x}$ est un zéro de $H(x)$.

Preuve : Le thm suit du lemme IV.3.3.5 de la même façon que le théorème IV.3.3.7 si on parvient à montré que Γ_a est borné. On prenons un $\delta > 0$ adapté au lemme IV.3.3.13. et supposons $\delta < r$. Alors, $\phi_a(\lambda, x) \neq 0$ à la surface $(\|x\|_\infty = r)$ de $Q = [0,1] \times \{x \mid x \geq 0; \|x\|_\infty \leq r\}$ par le lemme IV.3.3.13. Donc, par le lemme IV.3.3.11., $\Gamma_a \subset Q$ et sera par conséquent borné. ∎

Remarque :

On peut généraliser le théorème IV.3.3.14 en remplaçant $x_k F_k(x) > 0$ par $x_k > 0$ et $F_k(x) > 0$, et en supposant l'ensemble de solutions de (2) discret. Il en suit le théorème IV.3.3.15. La condition de non-singularité de $DH(x)$ aux racines de $H(\cdot)$ n'est pas nécessaire puisque l'ensemble de solutions de (2) est supposé discret.

Théorème IV.3.3.15.

Supposons que tout
zéro de $H(x)$ soit dans la boule $\|x\| < r$ où r est tel que $r > 0$ et $\|x\| \geq r$ entraîne $x_k > 0$ et $F_k(x) > 0$ pour un certain k.

Alors, il existe $\delta > 0$ tel que pour presque tout $a > 0$ avec $\|a\| < \delta$ il y a une courbe nulle Γ_a de $\phi_a(1,x)$ le long laquelle $D\phi_a(t,x)$ est de rang plein. Γ_a relie $(0,a)$ à $(1,\bar{x})$ où $\bar{x}$ est un zéro de $H(x)$.

Corollaire IV.3.3.16.

La conclusion du thm IV.3.3.14 tient au cas linéaire $F(x) = Mx + q$ où q est non-dégénéré par rapport à M et où M satisfait un des critères suivants : a) définie positive
b) une P-matrice
c) non-dégénérée strictement copositive.
d) non-dégénérée strictement semi-monotone.

Remarque: On essayera de se mettre aux conditions du thm IV.3.3.14

(i) La non-singularité de $DH(x)$ aux zéros de $H(x)$ suit de la non-dégénérescence de q par rapport M comme au corollaire IV.3.3.9.

(ii) Une matrice (PD) est une P-matrice. Une P-matrice est (SSM) par la remarque d à la page IV.23. En plus, il est évident qu'une matrice (SCP) est (SSM). Il suffit donc de traiter le cas d).

Définissons $\psi(x) \equiv \max_{x_i>0} (Mx)_i$, $x>0$. $\psi(\cdot)$ est continue et satisfait $\psi(\lambda x) = \lambda \cdot \psi(x)$ pour $\lambda>0$. Soit $\mu(r) = \min \psi(x)$ où $x \in \{ x \mid x \geq 0 ; \|x\|_\infty = r \}$, $r>0$. M est (SSM) ssi $\exists k$ tel que $x_k (Mx)_k > 0$ pour $x>0$. D'où, $\psi(x)>0$ et $\mu(r)>0$. En plus, $\mu(\lambda r) = \lambda \mu(r)$ pour $\lambda>0$ ce qui entraîne que $\mu(r) = r\mu(1)$, $r>0$ c-à-d $r = \frac{\mu(r)}{\mu(1)}$. Soit $r \geq 2 \frac{\|q\|_\infty}{\mu(1)} \equiv \tilde{r}$. Par conséquent, $\frac{\mu(r)}{\mu(1)} \geq 2 \frac{\|q\|_\infty}{\mu(1)}$ ce qui implique que $\mu(r) \geq 2\|q\|_\infty$ c-à-d que $\mu(r) > \|q\|_\infty$. Soit un $x>0$ tel que $\|x\|_\infty \geq \tilde{r}$. Il y a un indice k vérifiant $(Mx)_k = \psi(x) \geq \mu(\|x\|_\infty) > \|q\|_\infty \geq |q_k|$ ce qui entraîne que pour $x>0$ tel que $\|x\|_\infty \geq \tilde{r}$, il existe un k tel que $x_k (Mx+q)_k > 0$. ■

3) Le problème non-linéaire de complémentarité

Définition IV.3.3.17.

Soit $f : \mathbb{R}^n \to \mathbb{R}^n$ de classe C^3. Alors f est dit <u>uniformément convexe</u> ssi il existe $\nu>0$ tel que $x' A(z) x \geq \nu \|x\|^2$ pour tout $z, x \in \mathbb{R}^n$ où $A(z)$ est le Hessien de f en z.

IV. 4. Minimisation avec contraintes

IV. 4. 1. Position du problème

Nous traitons maintenant le problème de minimisation suivant :

$$(5)\quad \begin{cases} \min f(x) \\ g(x) \leq 0 \\ x \geqslant 0 \end{cases}$$

où $f : \mathbb{R}^n \to \mathbb{R}$ de classe C^3.
$g : \mathbb{R}^n \to \mathbb{R}^m$ de classe C^3.

On définit le lagrangien $L(x,u) \equiv f(x) + u\,g(x)$ où $u \in \mathbb{R}^m$.

Par les conditions de Kuhn-Tucker, nous savons que (5) admet une solution optimale $\bar{x}$ si et seulement si le système (6) possède une solution $(\bar{x},\bar{u}) \in \mathbb{R}^{n+m}$:

$$(6)\quad \begin{cases} (\bar{x},\bar{u}) \geqslant 0 \\ \nabla_x L(\bar{x},\bar{u}) \geqslant 0 \\ -\nabla_u L(\bar{x},\bar{u}) \geqslant 0 \\ \bar{x}^T \cdot \nabla_x L(\bar{x},\bar{u}) = 0 \\ -\bar{u}^T \nabla_u L(\bar{x},\bar{u}) = 0 \end{cases}$$

où $\nabla_x L : \mathbb{R}^{n+m} \to \mathbb{R}^n$
$\nabla_u L : \mathbb{R}^{n+m} \to \mathbb{R}^m$

Par définition $x^* \equiv (\bar{x},\bar{u})$ et $F(x^*) \equiv \begin{pmatrix} \nabla_x L(\bar{x},\bar{u}) \\ -\nabla_u L(\bar{x},\bar{u}) \end{pmatrix}$ où $F : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ (7)

On ramène donc de cette façon le système (6) au problème de complémentarité

$$(2)\quad \begin{cases} x^* \geqslant 0 \\ F(x^*) \geqslant 0 \\ x^* F(x^*) = 0 \end{cases}$$

Désignons dans la suite x^* par x.

Résoudre (2) revient (par IV.2.2.) à trouver la solution du système d'équations non-linéaires

$$(3) \qquad -(|F_i(x) - x_i|)^3 + (F_i(x))^3 + x_i^3 = 0 \qquad i = 1 \dots M+m$$

c-à-d $\qquad H(x) = 0$

IV. 4.2. Convergence de l'algorithme

On cherche la solution de (3) au moyen de l'application homotopique $\phi_a(\lambda, x) = \lambda H(x) + (1-\lambda)(x-a)$. Les réflexions dans IV.4.1. nous ont ramené aux conditions du paragraphe IV.3.3. L'algorithme convergera par conséquent si on satisfait les conditions du Thm IV.3.3. Il faut donc que F définit par (7) vérifie que l'ensemble des solutions de (2) soit discret et que tout zéro de $H(x)$ se trouve dans la boule $\|x\| < r$ où r est tel que $x \geq 0$ et $\|x\| \geq r$ entraîne $x_k > 0$ et $F_k(x) > 0$ pour un certain k.

Un point $X = (\bar{x}, \bar{u})$ ainsi trouvé nous donne $\bar{x} \in \mathbb{R}^n$ qui est solution optimale de (5).

On observe en pratique que les conditions sur F sont en général très difficiles à satisfaire pour des systèmes (6) quelconques.

IV. 5. Résultats numériques

IV.5.1. : Soit $\min_{x \in \mathbb{R}^n} f(x) = f(\bar{x})$ ie $\nabla f(\bar{x}) = 0$ (i)

où $f(x) = \sum_{i=1}^{n} x_i^4 - \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} i$

Comme ce minimum est borné, on peut utiliser (par le théorème III.1.1) l'application homotopique $\phi_a(\lambda, x) = \lambda \nabla f(x) + (1-\lambda)(x-a)$ qui trouvera en zéro de $\nabla f(x)$ c-à-d un minimum de f.

Dans les exemples IV.5.1 à III.5.4 on travaillera avec des bornes d'erreur EPS = 10.E-08, ARCTOL = 10.E-04 et avec $a = 0.0$.

Comme solution on trouve $x = (x_i)_{i=1\ldots n} = (0.88465\,E{+}00)_{i=1\ldots n}$

avec

N	NFE	ARCLEN
1	54	0.14392 E+01
2	58	0.18398 E+01
5	70	0.26426 E+01
10	80	0.35163 E+01
20	89	0.47182 E+01

IV.5.2. : Considérons $\min_{x \geq 0} f(x) = \min_{u \geq 0} \left\{ \sum_{i=1}^{n} x_i^2 + \sum_{\substack{i,j=1 \\ i<j}}^{n} 2 u_i u_j - \sum_{i=1}^{n} u_i \right.$ (1)

On est donc ramené à résoudre le problème linéaire de complémentarité

$$\begin{cases} x \geq 0 \\ \nabla f(x) \geq 0 \\ x \cdot \nabla f(x) = 0 \end{cases} \quad (2)$$

où $F(x) = Mx+q$ avec $M = \begin{pmatrix} 1 & 2 & 2 & \dots & 2 \\ 1 & 2 & & & 2 \\ 0 & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & 2 \\ & & & 0 & 1 \end{pmatrix}$ et $q = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}$

$= \nabla f(x)$

M est copositive, semi-définie positive et une P-matrice. Elle sont de très bonnes propriétés pour le problème linéaire de complémentarité. Par le corollaire IV.3.3.16. on trouve un zéro de $Mx+q$ qui est le minimum de $f(x) = x^T Mx + qx$. En annulant $H(x)$ défini en (3) on trouve le minimum $x = (x_i)_{i=1\dots n}$ où $\begin{cases} x(i) = 0 & \text{pour } i = 1 \dots n-1 \\ x(n) = 0.50000E+00 \end{cases}$

avec

N	NFE	ARCLEN
1	39	0.11321E+01
2	34	0.11710E+01
5	46	0.12059E+01
10	51	0.12256E+01
20	54	0.12401E+01
30	49	0.12480E+01
40	49	0.12524E+01
50	49	0.12558E+01

IV.5.3. Cherchons $\min_{x \geq 0} f(x) = \min_{x \geq 0} \left\{ \sum_{i=1}^{m} x_i^2 + 4 \sum_{\substack{i,j=1 \\ i \neq j}}^{m} x_i x_j - 2 \sum_{i=1}^{m} x_i \right\}$

Ceci revient à résoudre le problème linéaire de complémentarité (2)

$$\begin{cases} x \geq 0 \\ Mx+q \geq 0 \\ x(Mx+q) = 0 \end{cases} \qquad (2) \qquad \text{où } M = \begin{pmatrix} 1 & 4 & \dots & 4 \\ 4 & 1 & 4 & \vdots \\ 4 & 4 & \ddots & 4 \\ \vdots & & \ddots & 4 \\ 4 & 4 & \dots & 4 & 1 \end{pmatrix} \text{ et } q = \begin{pmatrix} -2 \\ \vdots \\ -2 \end{pmatrix}$$

Comme H est non-dégénérée et strictement semi-monotone on peut de nouveau appliquer le corollaire IV. 3. 3. 16.

Chaque minimum trouvé a toutes ses composantes égales.

Voici les résultats :

N	X = (Xi)i=1...M	NFE	ARCLEN
1	0.10000 E+01	49	0.14412 E+01
2	0.20000 E+00	52	0.11018 E+01
5	0.58824 E-01	55	0.10561 E+01
10	0.27027 E-01	51	0.10416 E+01
20	0.12987 E-01	51	0.10320 E+01
30	0.85470 E-02	103	0.10272 E+01
40	0.63694 E-02	98	0.10243 E+01
50	0.50761 E-02	98	0.10223 E+01

IV.5.4. Soit $\min_{x\in\mathbb{R}^n} f(x) = \min_{x\in\mathbb{R}^n} \left\{ \exp\left(\sum_{i=1}^{M} (x_i - i + 2)^2 \right) \right\}$

C'est le problème d'une fonction convexe à minimum borné. En appliquant le thm IV. 1. 1. on trouve la racine de f en annulant son gradient $\nabla f(x)$ au moyen de $\phi_a(b,x) = b\nabla f(x) + (1-b)(x-a)$.

La solution trouvée est $X = (-1., 0., 1., 2., \ldots, M-2)$ avec

N	NFE	ARCLEN
1	73	0.14626 E+01
2	69	0.14626 E+01
3	75	0.18402 E+01
4	85	0.27603 E+01

5	125	0.41501E+01
6	156	0.58436E+01

Pour $N > 6$, on n'obtient plus de résultats précis. Ceci est dû au fait que $\exp\left(\sum_{i=1}^{m}(x_i - i + 2)^2\right)$ dépasse la capacité de l'ordinateur. Une série "d'overflow" est la conséquence évidente.

Si on cherche $\min_{x \geq 0} f(x)$, on doit résoudre un problème non-linéaire de complémentarité ce qui peut se faire en appliquant l'algorithme cherchant les racines de (3) en se basant sur le thm IV.3.3.18.
Ainsi on trouve par exemple (0., 0., 1., 2., 3.) comme solution pour $n = 5$ avec NFE = 377 et ARCLEN = 0.45374 E+01.

IV.5.5. Soit
$$\min f(x) = (x_1 - 2)^2 + (x_2 - 1)^2$$
$$\text{SC} \begin{cases} g_1(x) = x_1^2 - x_2 \leq 0 \\ g_2(x) = x_1 + x_2 - 2 \leq 0 \end{cases} \qquad (1)$$

Pour se mettre aux conditions du paragraphe IV.4., définissons le lagrangien
$$L(x,u) = (x_1 - 2)^2 + (x_2 - 1)^2 + u_1(x_1^2 - x_2) + u_2(x_1 + x_2 - 2)$$

Posons $x_3 = u_1$, $x_4 = u_2$ et $x^* = (x_1, \ldots, x_4)$

$$F(x^*) = \begin{pmatrix} \nabla_x L(x,u) \\ -\nabla_u L(x,u) \end{pmatrix}$$

D'où : $F(x^*) = \begin{pmatrix} 2(x_1-2) + 2x_1x_3 + x_4 \\ 2(x_2-1) - x_3 + x_4 \\ -(x_1^2 - x_2) \\ -(x_1+x_2-2) \end{pmatrix}$

Il nous reste à trouver $x^* = (\bar{x}, \bar{u})$ tel que

$$-(|F_i(x^*) - x_i|)^3 + (F_i(x^*))^3 + x_i^3 = 0 \qquad i = 1 \dots 4$$

On trouve la solution $\bar{x} = (1,1)$

avec NFE = 286

ARCLEN = 0.14307 E+01

EPS = 0.10000 E-07

ARCTOL = 0.10000 E-02

Remarque : Notez que des problèmes de minimisation avec contraintes quelconques sont généralement impossible à résoudre puisque le système d'équations non-linéaires (3) est le plus souvent mal conditionné de façon à ce que les matrices résultantes soient trop compliquées et qu'on ne satisfait pas les conditions de convergence du théorème III.3.3.15. Pour ce genre de problèmes on conseille d'autres méthodes.

CHAPITRE V :

Problèmes à deux conditions limites

Une application très intéressante de l'algorithme trouvé est la résolution de problèmes à deux conditions limites (L.T. Watson, [12]).

V.1. Position du problème

Nous allons résoudre le problème de la forme suivante :

$$\begin{cases} \ddot{y}(t) = g(t, y(t), \dot{y}(t)) & 0 \le t \le 1 \qquad (1) \\ y(0) = 0 \quad y^*(1) = 0 \quad \text{où } y^* = y \text{ ou } y^* = \dot{y} & (2) \end{cases}$$

où $y(t) = (y_1(t), \dots, y_N(t))$ est un vecteur de dimension N.

$g(\cdot) : \mathbb{R}^{2N+1} \longrightarrow \mathbb{R}^N$ satisfait les conditions suivantes

a) $g(t,u,v)$ satisfait une condition lipschitzienne en (u,v) pour $0 \le t \le 1$ c-à-d

Pour tout $t \in [0,1]$; (u,v), $(u',v') \in \mathbb{R}^{2N}$

$$\| g(t,u(t),v(t)) - g(t,u'(t),v'(t)) \|_{\mathbb{R}^N} \le L \| (u(t),v(t)) - (u'(t),v'(t)) \|_{\mathbb{R}^{2N}}$$

b) g admet des dérivées partielles du second ordre continues par rapport à u et v pour $0 \le t \le 1$.

Les conditions a) et b) assurent l'existence d'une solution unique du problème (1)-(2).

Soit $u(t)$ la solution unique du problème à valeur initiale

$$\begin{cases} \ddot{y}(t) = g(t, y(t), \dot{y}(t)) \\ y(0) = 0 \ , \ \dot{y}(0) = x \end{cases} \qquad (3)$$

Définissons $f : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ par $f(x) = u^*(1)$. Les conditions sur g assurent que $f(\cdot)$ est une fonction de classe C^2. Le problème aux limites (1)-(2) est alors équivalent à

$$f(x) = 0 \qquad (4)$$

qu'on peut résoudre au moyen de l'application homotopique (5).

V.2. Théorèmes de convergence

Lemme V.2.1.

Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ une application de classe C^2 et définissons

$\phi_a : [0,1) \times \mathbb{R}^n \longrightarrow \mathbb{R}^m$ par $\phi_a(\lambda, x) = \lambda f(x) + (1-\lambda)(x-a)$ (5)

Alors, pour presque tout $a \in E^n$, il y a une courbe nulle γ de ϕ_a partant en $(0, a)$ suivant laquelle $D\phi_a(\lambda, x)$ est de rang plein.

Le lemme est une conséquence du théorème paramétrisé de Lord et du théorème II.3.2. La condition sur le rang implique que la courbe ne peut pas s'arrêter à l'intérieur de $[0,1) \times \mathbb{R}^n$ (cfr. dém. du thm III.3.2). On dit parfois que f est en même temps ouvert et fermé dans $[0,1) \times \mathbb{R}^n$.

Lemme V.2.2.

Si la courbe nulle f du lemme III.2.1. est bornée, elle admet un point d'accumulation $(1, \bar{x})$ où $f(\bar{x}) = 0$. En plus, si $Df(\bar{x})$ est non-singulière, alors f sera de longueur finie.

Preuve: Comme f est bornée, elle sera contenue dans le cylindre $[0,1] \times K$ où K est une boule fermée. Soit $0 < \alpha < 1$ et prenons un point $(\hat{\lambda}, \hat{x})$ quelconque de f appartenant à $[0, 1-\alpha] \times K$. Par le lemme V.2.1. $D\phi_a(\hat{\lambda}, \hat{x})$ est de rang plein, d'où par le thm des fonctions implicites f est de longueur finie au voisinage de $(\hat{\lambda}, \hat{x})$. Comme $[0, 1-\alpha] \times K$ est compact, la partie de f contenue dans $[0, 1-\alpha] \times K$ sera donc de longueur finie. Le lemme V.2.1. implique que f ne peut s'arrêter à l'intérieur de $[0, 1-\alpha] \times K$. Donc, f doit sortir de $[0, 1-\alpha] \times K$. Comme α est arbitraire, f admet un point d'accumulation dans $(1, \bar{x}) \in \{1\} \times K$. Puisque $\phi_a(1, \bar{x}) = 0$, on a par continuité que $f(\bar{x}) =$

Prolongeons $\phi_a(\lambda, x)$ sur $[0, 1+\alpha] \times \mathbb{R}^n$. Comme $Df(\bar{x})$ est non-singulière $D\phi_a(1, \bar{x}) = [f(x) - \bar{x} + a, Df(\bar{x})]$ sera de rang plein ce qui entraîne par le thm des fonctions implicites que f est de longueur finie dans un

voisinage de $(1,\bar{x})$ et passe par $(1,\bar{x})$. Il en suit que γ est de longueur finie dans $[0,1]\times K$. ■

La convergence de l'algorithme sera assurée par le théorème suivant

Théorème V.2.3.

Supposons qu'il existe une constante $M>0$ telle que $\|g(t,x(t),\dot{x}(t)\|_\infty \leq$ suivant toute trajectoire de (1) pour laquelle $x(0)=0$ et $\|\dot{x}(0)\|_\infty = M$.

Alors, pour presque tout $x\in\mathbb{R}^n$ avec $\|a\|_\infty < M$, il y a une courbe nulle γ de $\phi_a(1,x)$ le long de laquelle $D\phi_a(1,x)$ est de rang plein, se trouvant dans $[0,1]\times\{x\in\mathbb{R}^n \mid \|x\|_\infty < M\}$ et liant $(0,a)$ à $(1,\bar{x})$ où $\bar{x}$ est un zéro de f. Si $Df(\bar{x})$ est non-singulière, γ sera de longueur finie.

Avant de voir la preuve de ce théorème fondamental, nous allons encore étudier le

Lemme V.2.4.

Sous les hypothèses du théorème, $\phi_a(1,x)$ n'est jamais nulle dans $[0,1)\times\{x\in\mathbb{R}^n \mid \|x\|_\infty = M\}$.

Preuve : Soit $u(t)$ solution du problème à valeur initiale

$$\begin{cases} \ddot{y}(t) = g(t,y(t),\dot{y}(t)) \\ y(0)=0 \quad , \quad \dot{y}(0)=x \end{cases} \qquad \text{où } \|x\|_\infty = M$$

Supposons $f(x) \equiv u^*(1) = \dot{u}(1)$. Alors, $\|f(x) - x\|_\infty \triangleq \|\dot{u}(1) - \dot{u}(0)\|_\infty$

$= \|\int_0^1 \ddot{u}(t)\,dt\| = \|\int_0^1 g(t, u(t), \dot{u}(t))\,dt\| \leq h \cdot 1 = h$ par hypothèse.

Supposons $f(x) \equiv u^*(1) = u(1)$. Alors, $\|f(x) - x\|_\infty = \|u(1) - \dot{u}(0)\|_\infty$

$= \|(u(1) - u(0)) - \dot{u}(0)\|_\infty = \|\int_0^1 (\dot{u}(\tau) - \dot{u}(0))\,d\tau\|_\infty$

$= \|\int_0^1 \int_0^\tau g(s, u(s), \dot{u}(s))\,ds\,d\tau\|_\infty \leq \|\int_0^1 \tau \cdot h \cdot d\tau\|_\infty = \frac{\tau^2}{2}\Big]_0^1 M = \frac{h}{2} < h$

Dans les deux situations possibles $f(x)$ se trouve dans la boule de rayon h et de centre x. Comme $\|a\|_\infty < h$, $x - a$ se trouve dans la même boule. Il est maintenant évident que $f(x)$ et $x - a$ ne peuvent pointer dans des directions opposées de façon à ce que $\phi_a(\lambda, x) = \lambda f(x) + (1-\lambda)(x-a) \neq 0 \quad ; 0 \leq \lambda \leq 1$. ∎

Preuve du théorème V.2.3.

L'existence de la courbe γ, suivant laquelle $D\phi_a(\lambda, x)$ est de plein rang, partant de $(0, a)$ pour presque tout a où $\|a\|_\infty < h$ suit du lemme V.2.1. Le fait que γ soit contenue dans le cylindre $[0,1] \times \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq h\}$ suit du lemme V.2.4. γ est donc bornée et le lemme V.2.2 entraîne qu'on atteint $(1, \bar{x})$ où $\bar{x}$ est un zéro de f et que γ est de longueur finie si $Df(\bar{x})$ est non-singulière. ∎

Le théorème V.2.3. nous permet maintenant de résoudre (1)-(2).

V.3. L'algorithme

Le problème revient à utiliser l'algorithme décrit au chapitre III appliqué à l'application homotopique $\Phi_a(\lambda, x) = \lambda f(x) + (1-\lambda)(x-a)$. Cela équivaut (après paramétrisation par la longueur d'arc) à résoudre

$$\begin{cases} \frac{d}{ds}\Phi_a(\lambda(s), x(s)) = 0 & \lambda(0) = 0 \\ \left\| \left(\frac{d\lambda}{ds}, \frac{dx}{ds}\right) \right\|_2 = 1 & x(0) = a \end{cases} \qquad (5)$$

On travaillera par conséquent avec l'algorithme du point fixe adapté aux recherches de racines (voir III.3.3).

Pour définir la fonction $f(x)$ à partir de (1)-(2), il suffit d'intégrer le système (3) au moyen d'une méthode de Runge-Kutta à 4 étapes par exemple ou d'utiliser d'autres intégrations numériques.

V.4. Résultats numériques

L'exemple suivant satisfait aux hypothèses du théorème V.2.3.

$$\begin{cases} \ddot{y_1} = (1 + y_1'^2 + y_2^2 + y_3^2)^{-1} \\ \ddot{y_2} = (2 + y_1^2 + y_2'^2 + y_3^2)^{-1} \\ \ddot{y_3} = (3 + y_1^2 + y_2^2 + y_3'^2)^{-1} \end{cases} \qquad (6)$$

$$y(0) = y(1) = 0$$

Soit $z_1 = y_1$; $z_2 = y_2$; $z_3 = y_3$; $z_4 = \dot{y_1}$; $z_5 = \dot{y_2}$; $z_6 = \dot{y_3}$

Le système (6) aura maintenant la forme

$$(7)\quad \begin{cases} z_1' = z_4 \\ z_2' = z_5 \\ z_3' = z_6 \\ z_4' = (1 + z_4^2 + z_2^2 + z_3^2)^{-1} \\ z_5' = (2 + z_1^2 + z_5^2 + z_3^2)^{-1} \\ z_6' = (3 + z_1^2 + z_2^2 + z_6^2)^{-1} \end{cases} \qquad SC \begin{cases} z_1(0) = z_2(0) = z_3(0) = 0 \quad (8) \\ z_1(1) = z_2(1) = z_3(1) = 0 \quad (9) \end{cases}$$

Au moyen de la méthode de Runge-Kutta à 4 étages (le pas $h = 0.0$

on intégre le système

$$(7)\quad SC \begin{cases} (8) \\ z_4(0) = x_1 \; ; z_5(0) = x_2 \; ; z_6(0) = x_3 \end{cases}$$

ce qui nous fournit un système équivalent à (6) de la forme

$$x = \begin{pmatrix} z_4(0) \\ z_5(0) \\ z_6(0) \end{pmatrix} \quad \text{et} \quad f(x) = \begin{pmatrix} z_1(1) \\ z_2(1) \\ z_3(1) \end{pmatrix}$$

où on cherche x tel que $f(x) = 0$. (4)

Nous trouvons les résultats suivants avec EPS = 1.0E-09

ARCTOL = 1.E-04

a = 0.0

NFE = 92

ARCLEN = 0.15342E+01

X = -0.46505E+00 -0.24645E+00 -0.16563E+00

CHAPITRE VI :

Polynômes à racines réelles

On appliquera dans la suite l'algorithme du point fixe aux polynômes à racines réelles.

VI.1. Position du problème

Considérons la classe de polynômes moniques de degré n quelconques du corps réel sur lui-même ayant toutes les racines réelles :

$$P(x) = x^n + a_n x^{n-1} + \dots + a_2 x + a_1$$

Pour chercher ces racines il suffit de ne considérer les polynômes en question que sur un compact $[b,c]$, $b<0$, $c>0$ contenant toutes les racines et qui est tel que

$$\overline{P}(\cdot) : [b,c] \longrightarrow [b,c]$$

où $\overline{P}(x) = x^n + a_n x^{n-1} + \dots + (a_2+1)x + a_1$

Cette classe de polynômes aura donc l'allure de la figure 1.

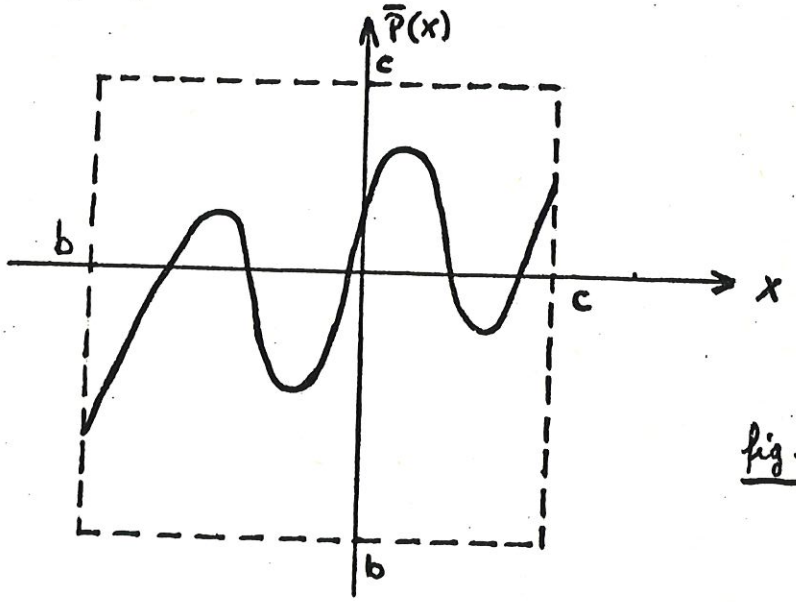


fig. 1

Le compact $[b,c]$ jouera le rôle de la boule unité dans le chapitre II. Comme un polynôme réel est une application régulière, on peut utiliser le théorème II.3.2. qui assure l'existence d'une courbe nulle de $\Phi_a(\lambda,x) = \lambda(x - \overline{P}(x)) + (1-\lambda)(x-a)$ et qui trouve un point fixe de $\overline{P}(x)$ c-à-d une racine de $P(x)$ notée x_1. En factorisant $P(x)$ par rapport à $(x-x_1)$ au moyen du schéma de Hörner, on trouvera $P_1(x)$ tel que $P(x) = P_1(x)(x-x_1)$. En continuant ce raisonnement sur $P_1(x)$ de même façon que sur $P(x)$

on tombera sur x_2 racine de $P_1(x)$ (d'où aussi de $P(x)$). On trouve alors par Hörner un $P_2(x)$ tel que $P(x) = P_1(x)(x-x_2) = P_2(x)(x-x_1)(x-x_2)$

On continuera ce raisonnement jusqu'à ce que toutes les racines sont trouvées.

Si $P(x)$ est à m ($m<n$) racines réelles, l'algorithme appliqué à $P_m(x)$ divergera puisqu'il n'y a plus de points fixes de $\bar{P}_m(x)$

VI.2. Algorithme numérique

On utilise l'algorithme décrit au chapitre III adapté à l'application homotopique $\Phi_a(\lambda, x) = \lambda(x - \bar{P}(x)) + (1-\lambda)(x-a)$ où $\bar{P}(x) = P(x) + x$

Il faut procéder de la façon suivante :

#1 : Appliquer l'algorithme du point fixe à $\Phi_a(\lambda, x)$
ce qui me donne la racine $\bar{x}$ de $P(x)$.

#2 : Utiliser une factorisation de Hörner de $P(x)$ par rapport à $(x-\bar{x})$; D'où, on trouve $P_1(x)$ tel que $P(x) = P_1(x)(x-\bar{x})$.
Poser $P(x) = P_1(x)$.
Si degré de $P(x) > 1$, aller en #1, sinon $P(x) = x + a_1$ et la dernière racine $\bar{x}$ vaut $-a_1$, STOP.

VI.3. Résultats numériques

Travaillons dans la suite avec EPS= 0.10000E-06

ARCTOL= 0.10000 E-02 .

VI.3.1. Considérons $P(x) = x^4 + 0.42781E-02\ x^3 - 0.12373E+01\ x^2 - 0.23410E-02 \cdot x + 0.25000E+00$

→ Le premier appel de l'algorithme du point fixe donne la racine 0.50345E+00 avec NFE = 49 et ARCLEN= 0.11250E+01.

Une factorisation de Hörner donne $P(x) = P_1(x) \cdot (x - 0.50345)$

où $P_1(x) = x^3 + 0.50733E+00\ x^2 - 0.98168E+00 \cdot x - 0.49657E+0$

Continuer avec "P(x)" = $P_1(x)$.

→ Le deuxième appel donne -0.50517E+00 avec NFE=70 et ARCLEN=0.11211E+

Hörner fournit $P_1(x) = x^2 + 0.25585E-02\ x - 0.98297E+00$

→ Troisième appel : -0.99273E+00 avec NFE= 89 et ARCLEN= 0.14230E+01

$P_1(x) = x - 0.99017E+00$.

→ La dernière racine vaut 0.99017E+00.

On obtient donc les racines suivantes :

0.50345 -0.50517 -0.99273 0.99017 .

VI.3.2. Soit $P(x) = x^4 + 3x^3 + x^2 - 2x - 1$

→ -0.55496 avec NFE = 43 et ARCLEN = 0.11504E+01.

$P_1(x) = x^3 + 2.445\,x^2 - 0.3569\,x - 1.8019$.

→ -1.0 avec NFE = 45 et ARCLEN = 0.14512 E+01

$P_2(x) = x^2 + 1.445\,x - 1.8019$

→ -2.2470 avec NFE = 66 et ARCLEN = 0.25050 E+01

$P_3(x) = x - 0.80194$

→ 0.80194.

VI.3.3. Considérons $P(x) = x^6 + 7x^5 + 7x^4 - 35x^4 - 56x^2 + 28x + 48$

→ 1.0 avec NFE = 63 et ARCLEN = 0.18654 E+01.

$P_1(x) = x^5 + 8x^4 + 15x^3 - 20x^2 - 76x - 48$.

→ -1.0 avec NFE = 50 et ARCLEN = 0.17242 E+01.

$P_2(x) = x^4 + 7x^3 + 8x^2 - 28x - 48$.

→ -2.0 avec NFE = 64 et ARCLEN = 0.25073 E+01.

$P_3(x) = x^3 + 5x^2 - 2x - 24$

→ -3.0 avec NFE = 74 et ARCLEN = 0.33328 E+01.

$P_4(x) = x^2 + 4 - 8$

→ -4.0 avec NFE = 87 et ARCLEN = 0.42338 E+01.

$P_5(x) = x - 2$.

→ 2.0

ANNEXE :

① Les applications uniformément convexes

A.1. Définitions

Soit $D_0 \subset \mathbb{R}^n$ une partie convexe ; $D_0 \subset D$.

Une fonction $g : D_0 \to \mathbb{R}$ est dite convexe

ssi pour tout $x, y \in D_0$ et $0 < \alpha < 1$

$$g(\alpha x + (1-\alpha) y) \leq \alpha\, g(x) + (1-\alpha)\, g(y) \qquad (1)$$

La fonction g est strictement convexe

ssi l'inégalité (1) tient de façon stricte si $x \neq y$.

La fonction g est uniformément convexe

ssi il existe une constante $c > 0$ telle que pour $x, y \in D_0$ et

$0 < \alpha < 1$ $\quad \alpha\, g(x) + (1-\alpha) g(y) - g(\alpha x + (1-\alpha) y) \geq c\, \alpha(1-\alpha)\, \|x-y\|^2$ (2

On essayera dans la suite d'établir l'équivalence entre la définition (2) et celle donnée au chapitre III (IV 3.3.17.)

Un exemple typique de fonction convexes est $g(x) = x'Ax$ où A est une matrice semi-définie positive.

Il est évident qu'une fonction convexe dérivable se trouve toujours au-dessus de l'hyperplan en chacun de ses points c-à-d

$$\forall x, y \in D_0 \quad g'(x)(y-x) \leq g(y) - g(x) \qquad (3)$$

ce qui me donne qu'il existe $c > 0$ tel que

$$\forall x, y \in D_0 \quad g'(x)(y-x) - (g(y)-g(x)) \leq c\|x-y\|^2 \qquad (3')$$

dans le cas des fonctions uniformément convexes.

A.2. Lemme

Supposons $g : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ de classe C^1 sur un convexe $D_0 \subset D$.

Alors, g est convexe sur D_0 ssi $(g'(y)-g'(x))(y-x) \geq 0 \quad \forall x, y \in D_0$ (4)

et g est strictement convexe ssi l'inégalité (4) tient de façon stricte si $x \neq y$. Finalement, g est uniformément convexe sur D_0 ssi $(g'(y)-g'(x))(y-x) \geqslant 2c\|y-x\|^2 \quad \forall x, y \in D_0$ (5)

où $c > 0$ est la constante de (2).

Preuve: Si g est uniformément convexe sur D_0, alors $\forall x, y \in D_0$ on a

$$\begin{aligned} g(y) - g(x) &\geqslant g'(x)(y-x) + c\|y-x\|^2 \\ g(x) - g(y) &\geqslant g'(y)(x-y) + c\|x-y\|^2 \end{aligned} \qquad (6)$$

et dans le cas convexe (6) tient avec $c=0$. En additionnant les deux inégalités de (6), on obtient (5) et donc, (4) est vrai pour $c=0$ dans le cas convexe. En plus, si g est strictement convexe, alors (6) tient avec $c=0$, mais inégalité stricte puisque $x \neq y$ et d'où par addition, (4) est respecté avec l'inégalité stricte pour $x \neq y$.

Pour démontrer les conditions suffisantes remarquons que pour x, y dans D_0 fixés, le thm de la moyenne assure qu'il existe $t \in (0,1)$ tel que

$$g(y) - g(x) = g'(u)(y-x) \qquad (7)$$

où $u = x + t(y-x)$. On déduit par (4) que

$$(g'(u) - g'(x))(y-x) = \left(\frac{1}{t}\right)(g'(u) - g'(x))(u-x) \geq 0 \qquad (8)$$

et donc

$$\begin{aligned} g(y) - g(x) &= (g'(u) - g'(x))(y-x) + g'(x)(y-x) \\ &\geq g'(x)(y-x) \qquad (9) \end{aligned}$$

ce qui montre que g est convexe. Si l'inégalité est stricte dans (4) pour $x \neq y$, alors il en sera de même dans (8) ce qui implique la convexité stricte de g.

Finalement, si (5) est satisfait, soit $t_k = \frac{k}{m+1}$, $k = 0, 1, \ldots m+1$ où m est un entier arbitraire positif. Par le théorème de la

moyenne, il existe des nombres s_k tels que

$$g(x + t_{k+1}(y-x)) - g(x + t_k(y-x))$$

$$= g'(x + s_k(y-x))\ (t_{k+1} - t_k)(y-x) \quad \text{où } t_k < s_k < t_{k+1}$$

D'où, $g(y) - g(x) = \sum_{k=0}^{m} g(x + t_{k+1}(y-x)) - g(x + t_k(y-x))$

$$= \sum_{k=0}^{m} [g'(x + s_k(y-x)) - g'(x)]\ (t_{k+1} - t_k)(y-x) + \sum_{k=0}^{m} g'(x)(y-x)(t_{k+1} - t_k)$$

$$\geq 2c \|y-x\|^2 \sum_{k=0}^{m} (t_{k+1} - t_k)\, s_k + g'(x)(y-x)$$

Mais, $\sum_{k=0}^{m} (t_{k+1} - t_k)\, s_k \geq \sum_{k=0}^{m} (t_{k+1} - t_k)\, t_k$

$$= \sum_{k=0}^{m} \frac{1}{m+1} \cdot \frac{k}{m+1} = \sum_{k=0}^{m} k \cdot \frac{1}{(m+1)^2} = \frac{1}{(m+1)^2} \cdot \frac{m+1 \cdot m}{2} = \frac{1}{2} \frac{m}{m+1} \geq \frac{1}{2}$$

Il en suit que $g(y) - g(x) \geq c\|y-x\|^2 + g'(x)(y-x)$ ce qui assure la convexité uniforme de g. ∎

Soit $H(x)$ le Hessien de g en x. Notons que la G-dérivée seconde g'' de g est <u>définie positive</u> en x si $g''(x)\, h\, h > 0$ pour tout $h \in \mathbb{R}^n \neq 0$. Elle est <u>semi-définie positive</u> en x si $g''(x)\, h\, h \geq 0$ $\forall h \in \mathbb{R}^n$ et <u>uniformément définie positive</u> sur un ensemble D_0 s'il existe une constante $c > 0$ telle que $g''(x)\, h\, h \geq c\|h\|^2$ pour tout $h \in \mathbb{R}^n$ et $x \in D_0$. Remarquons que ces définitions

n'exigent pas $g''(x)$ symétrique. On suppose que la seconde G-dérivée $g''(x)$ existe c-à-d par définition

$$\text{pour tout } h \in \mathbb{R}^n \quad \lim_{t \to 0} \frac{\| g'(x+th) - g'(x) - t\, g''(x)\, h \|}{t} = 0$$

En plus, $g''(x)$ est définie positive ssi $H(x)$ l'est.

A.3. Théorème

Supposons $g : D \subset \mathbb{R}^n \to \mathbb{R}$ admet une seconde G-dérivée en tout point d'un convexe $D_0 \subset D$.

Alors, g est convexe sur D_0 ssi $g''(x)$ est semi-déf. positive pour tout $x \in D_0$. En plus, g est strictement convexe sur D_0 si $g''(x)$ est définie positive en tout point x de D_0 et g est unif. convexe sur D_0 ssi g'' est unif. définie positive sur D_0.

Preuve: Soient $x, y \in D_0$. Alors par le thm de la moyenne, il existe un point $u \in [x, y]$ tel que

$$g(y) - g(x) - g'(x)(y-x) = \frac{1}{2} g''(u)(y-x)(y-x) \qquad (10)$$

D'où, g'' semi-déf. positive, définie positive ou unif. définie positive sur D_0 implique par le lemme que g est convexe, strictement convexe et uniformément convexe sur D_0.

Inversément, si g est uniformément convexe, le lemme montre que

$$g''(x)\,h\,h = \lim_{t\to 0}\left(\frac{1}{t}\right)\left(g'(x+th)-g'(x)\right)h$$

$$\geq \lim_{t\to 0}\left(\frac{1}{t^2}\right) 2c\,\|th\|^2 = 2c\,\|h\|^2 \quad \forall x\in D_0,\ \forall h\in\mathbb{R}^n. \tag{11}$$

où $c>0$ est la constante de (2).

Par conséquent g'' est unif. définie positive sur D_0. Finalement si g est convexe sur D_0, le lemme assure de nouveau que (11) tient avec $c=0$, donc g'' est semi-définie positive. ∎

Le thm ne dit pas que la convexité stricte implique que g'' soit définie positive. En fait, la fonction x^4 est strictement convexe, mais admet une dérivée seconde nulle en $x=0$.

Donc, par le thm A.3. on peut conclure que
g est <u>uniformément convexe</u> ssi
il existe $c>0$ tel que $x'H(z)x \geq c\,\|x\|^2$ pour tout $x, z\in\mathbb{R}^n$.

Ceci établit l'équivalence entre la définition IV.3.3.17. et celle donnée au début de ce chapitre. Pour plus de détails consultez [19].

② La subroutine FIXPT

```
      SUBROUTINE FIXPT(N,Y,ARCTOL,EPS,ARCLEN,NFE,IFLAG)
C     ------------------------------------------------
C
C     DECLARATIONS
C     ------------
C
      REAL Y(101),WT(101),PHI(101,16),P(101),YP(101),PSI(12),YOS(100)
      REAL YPOLD(101),A(100)
      COMMON /FIXEDP/ YPOLD,A,NFEC,NC,NP1,IFLAGC
      EXTERNAL FODE
      LOGICAL START,CRASH,ST99
C     LIMITD EST UNE BORNE SUPERIEURE DU NOMBRE DE PAS. ON PEUT
C     LE CHANGER EN MODIFIANT LE DATA STATEMENT SUIVANT
      DATA LIMITD/1000/

C     SUIVANT LES VALEURS DE IFLAG ON DECIDE COMMENT IL FAUT SUIVRE
C     LA COURBE GAMMA.

      IF (IFLAG.EQ.0) GOTO 10
      IF (IFLAG.EQ.2) GOTO 35
      IF (IFLAG.EQ.3) GOTO 30
C     SEUL INPUT VALABLE POUR IFLAG =0,2,3
      RETURN


C     INITIALISATION
C     --------------

   10 ARCLEN=0.0
      S=0.0
      IF (ARCTOL.LE.0.0) ARCTOL=.5*SQRT(EPS)
      NFEC=0
      NC=N
      IFLAGC=0
      NP1=N+1
      SQNP1=SQRT(FLOAT(NP1))
      CURTOL=.01/SQNP1
      ST99=.FALSE.
      START=.TRUE.
      CRASH=.FALSE.
      H=.1
      EPSSTP=ARCTOL

C     METTRE DES CONDITIONDS INTIALES POUR L'ED ORDINAIRE
C     ---------------------------------------------------

      YPOLD(1)=1.0
      Y(1)=0.0
      WT(1)=1.0
      DO 20 J=2,NP1
      YPOLD(J)=0.0
      A(J-1)=0.0
      Y(J)=A(J-1)
      WT(J)=1.0
   20 CONTINUE
   30 LIMIT=LIMITD
```

```
C     FIN DU BLOCK D'INITIALISATION

C     DEPART DU RPOGRAMME
C     -------------------

   35 DO 150 ITER=1,LIMIT
      IF(Y(1).GE.0.0) GOTO 50
   40 IFLAG=5
      RETURN
   50 IF(S.LE.0.7*SQNP1) GOTO 80
C     LA LONGEUR D'ARC EST TROP GRANDE (D(LAMBDA)/DS TROP PETIT
C     ON REPART LE PROBLEME AVEC UNE AUTRE VALEUR DE A
      ARCLEN=ARCLEN+S
      S=0.0
   60 START=.TRUE.

      CRASH=.FALSE.
C     CALCULER UN NOUVEAU VECTEUR A
      DO 67 IUN=1,N
      YOS(IUN)=Y(IUN+1)
   67 CONTINUE
      CALL F(YOS,A,N)
      DO 70 JW=1,N
      A(JW)=(Y(JW+1)-Y(1)*A(JW))/(1.0-Y(1))
      IF (ABS(A(JW)).GT. 0.95) GOTO 40
   70 CONTINUE

      GOTO 100
   80 IF(Y(1).LE. .99 .OR. ST99) GOTO 100
C     SI LAMBDA ATTEINT .99 ON REPART LE PROBLEME AVEC UN NOUVEAU VECTEUR A

   90 ST99=.TRUE.
      EPSSTP=EPS
      ARCTOL=EPS
      GOTO 60

C     FAIRE UN PAS LE LONG DE LA COURBE.
  100 CALL STEP(S,Y,FODE,NP1,H,EPSSTP,WT,
     1    START,HOLD,K,KOLD,CRASH,PHI,P,YP,PSI)

      NFE=NFEC
C     VOIR SI LE PAS A EU DU SUCCES
      IF (IFLAGC.NE.4) GOTO 120
      IFLAG=4
      RETURN
  120 IF(.NOT. CRASH) GOTO 130
C     RETURN CODE SI LA TOLERANCE POUR L'ERREUR EST TROP PETITE
      IFLAG=2
C     CHANGER LES TOLERANCES D'ERREUR
      EPS=EPSSTP
      IF(ARCTOL.LT.EPS) ARCTOL=EPS
C     CHANGER LE NOMBRE LIMITE D'ITERATIONS
      LIMIT=LIMIT-ITER
      RETURN
  130 EPSSTP=ARCTOL
      IF(ABS(YP(1)).LE.CURTOL) EPSSTP=EPS
      IF(Y(1).LT.1.0) GOTO 150
      IF(ST99) GOTO 160
```

```
C     SI LAMBDA PLUS GRAND 1.0, MAIS QUE LE PROBLEME N'A PAS ETE
C     REPARTI AVEC UN NOUVEAU A, RETOURNER ET REPARTIR.
      S99=S-.5*HOLD

C     TROUVER UN ZERO Y(S) AVEC T(1)=LAMBDA.LT.1.0
  135 CALL INTRP(S,Y,S99,WT,P,NP1,KOLD,PHI,PSI)
      IF(WT(1).LT.1.0) GOTO 140
      S99=.5*(S-HOLD+S99)
      GOTO 135
  140 DO 144 JUDY=1,NP1
      Y(JUDY)=WT(JUDY)
      YPOLD(JUDY)=P(JUDY)
  144 WT(JUDY)=1.0
      S=S99
      GOTO 90
  150 CONTINUE
C     LAMBDA N'ATTEINT PAS 1 EN 1000 PAS
      IFLAG=3
      RETURN

C     CALCUL DU POINT FIXE
C     --------------------

C     UTILISER UNE INTERPOLATION INVERSE POUR TROUVER LA
C     REPONSE EN LAMBDA EGAL 1.0
  160 SA=S-HOLD
      SB=S
      LCODE=1

  170 CALL ROOT(SOUT,Y1SOUT,SA,SB,EPS,EPS,LCODE)
C     ROOT CHERCHE S TEL QUE Y(1)(S)=LAMBDA(S)=1.0
C     CETTE VALEUR SE TROUVERA DANS SA APRES L'APPEL DE ROOT.
      IF (LCODE.GT.0) GOTO 190
      CALL INTRP(S,Y,SOUT,WT,P,NP1,KOLD,PHI,PST)
      Y1SOUT=WT(1)-1.0
      GOTO 170

  190 IFLAG=1
C     METTRE IFLAG=6 SI ROOT NE TROUVE PAS LAMBDA =1.0
      IF (LCODE.GT.2) IFLAG=6
      ARCLEN=ARCLEN+SA
C     LAMBDA(SA)=1
C     ON CHERCHE LE POINT FIXE PAR INTERPOLATION POLYNOMIALE.
      CALL INTRP(S,Y,SA,WT,P,NP1,KOLD,PHI,PSI)
      DO 210 J=1,NP1
  210 Y(J)=WT(J)
      RETURN
      END
```

References

[1] : D.R. Smart : Fixed Point theorems
Cambridge University Press ; 1974.

[2] : Hilton and Wylie : Homology theory
Cambridge Uni. Press ; 1965

[3] : N. Bourbaki : Espaces vectoriels topologiques ; 5.1.4 ; 1955

[4] : Chow, York, Mallet : Finding zeros of maps : Homotopy methods that are constructive with probability one.
Mathematics of computation ; volume 32, N° 143
July 1978, Pages 887-899.

[5] : John W. Milnor : Topology from the differentiable Viewpoint
Univ. Press of Virginia ; Charlottesville.

[6] : Th. Bröcker : Differentiable Germs and Catastrophes
Univ. Regensburg (Math.)

[7] : H.L. Royden : Real Analysis
Mac Millan ; New York, 1968

[8] : L.T. Watson : A globally convergent algorithm for computing fixed points of C^2 maps.

Applied mathematics comput. (1979)

[9] : L.F. Shampine and M.K. Gordon : Computer Solution of ODE: The initial value problem.

W.H. Freeman ; San Francisco , 1975.

[10] : P. Businger : Linear least squares solution by Householder transformation.

Numerische Mathematik , 7 (1965) ; pages 269-276.

[11] : L.T. Watson : Fixed points of C^2 maps (Algorithm)

Virginia State University ; Blacksburg , USA ; 1979

[12] : L.T. Watson : The two point boundary value problem.

Dept of Math. , Michigan State University ,

Siam J. Numerical analysis , 1979.

[13] : L.T. Watson : Computational experience with the Chow - York algorithm.

Dept of Computer Science ; Virginia State Univ. , 1978

[14] : O.L. Mangasarian : Non-linear Programming

Mac Graw - Hill ; New - York , San Francisco , 1969

[15] : O.L. Mangasarian : Equivalence of the complementary problem to a system of non-lin. équations

Siam J, Numerical analysis, 13 (1976) ; pp 473-488.

[16] : L.T. Watson : Solving the non-linear complementary problem by a homotopy method.

Siam J, Control and Optimisation ; Vol 17, N°1, 1979

[17] : L.T. Watson : A variationnal approach to the linear complementary problem.

Doctoral dissertation Univ. of Michigan, 1974

[18] : M. Fiedler and V. Ptak : On matrices with non-positive off-diagonal elements and positive principal minors.

Czechoslovak Math., 1962, pp 473-483.

[19] : Ortega and Rheinboldt : Iterative solution of Nonlinear Equations in several variables

Table des matières

Chapitre IV : Applications en optimisation

Chapitre V : Problèmes à deux conditions limites